Iterative Feature Selection In Least Square Regression Estimation

Pierre Alquier

ITERATIVE FEATURE SELECTION IN LEAST SQUARE REGRESSION ESTIMATION

PIERRE ALQUIER

Abstract. In this paper, we focus on regression estimation in both the inductive and the transductive case. We assume that we are given a set of features (which can be a base of functions, but not necessarily). We begin by giving a deviation inequality on the risk of an estimator in every model defined by using a single feature. These models are too simple to be useful by themselves, but we then show how this result motivates an iterative algorithm that performs feature selection in order to build a suitable estimator. We prove that every selected feature actually improves the performance of the estimator. We give all the estimators and results at first in the inductive case, which requires the knowledge of the distribution of the design, and then in the transductive case, in which we do not need to know this distribution.

1. The setting of the problem

We give here notations and introduce the inductive and transductive settings.

1.1. Transductive and inductive settings.

Let $(\mathcal{X}, \mathcal{B})$ be a measure space and let $\mathcal{B}_{\mathbb{R}}$ denote the Borel σ-algebra on $\mathbb{R}$.

1.1.1. *The inductive setting.* In the inductive setting, we assume that P is a distribution on pairs $Z = (X, Y)$ taking values in $(\mathcal{X} \times \mathbb{R}, \mathcal{B} \otimes \mathcal{B}_{\mathbb{R}})$, that P is such that:

$$P\,|Y| < \infty,$$

and that we observe N independent pairs $Z_i = (X_i, Y_i)$ for $i \in \{1, ..., N\}$. Our objective is then to estimate the regression function on the basis of the observations.

Definition 1.1 (The regression function)**.** We denote:

$$\begin{aligned} f : \mathcal{X} &\to \mathbb{R} \\ x &\mapsto P(Y|X = x). \end{aligned}$$

1.1.2. *The transductive setting.* In the transductive case, we assume that P_{2N} is some exchangeable probability measure on the space $\left((\mathcal{X} \times \mathbb{R})^{2N}, (\mathcal{B} \otimes \mathcal{B}_{\mathbb{R}})^{\otimes 2N}\right)$. We will write $(X_i, Y_i)_{i=1...2N} = (Z_i)_{i=1...2N}$ a random vector distributed according to P_{2N}.

Definition 1.2 (Exchangeable probability distribution)**.** Let $\mathfrak{S}_k$ denote the set of all permutations of $\{1, ..., k\}$. We say that P_{2N} is exchangeable if for any $\sigma \in \mathfrak{S}_{2N}$ we have: $(X_{\sigma(i)}, Y_{\sigma(i)})_{i=1...2N}$ has the same distribution under P_{2N} that $(X_i, Y_i)_{i=1...2N}$.

Date: November 11, 2005.

2000 *Mathematics Subject Classification.* Primary 62G08; Secondary 62G15, 68T05.

Key words and phrases. Regression estimation, statistical learning, confidence regions, thresholding methods, support vector machines.

I Would like to thank my PhD advisor, Professor Olivier Catoni, for his constant help.

We assume that we observe $(X_i, Y_i)_{i=1...N}$ and $(X_i)_{i=N+1...2N}$; $(X_i, Y_i)_{i=1...N}$ is usually called the training sample and $(X_i, Y_i)_{i=N+1...2N}$ the test sample. In this case, we only focus on the estimation of the values $(Y_i)_{i=N+1...2N}$. This is why Vapnik [12] called this kind of inference "transductive inference".

Note that in this setting, the pairs (X_i, Y_i) are not necessarily independent, but are indentically distributed. We will let P denote their marginal distribution, and we can here again define the regression function f.

1.2. **The model.** In both settings, we are going to use the same model to estimate the regression function. Let Θ be a vector space, and:

$$\begin{aligned} F : \Theta \times \mathcal{X} &\to \mathbb{R} \\ (\theta, x) &\mapsto F(\theta, x) = f_\theta(x) \end{aligned}$$

be such that, for any $x_0 \in \mathcal{X}$, the application $\theta \mapsto f_\theta(x_0)$ is linear. We define the model:

$$\mathcal{F} = \{f_\theta(.), \theta \in \Theta\}.$$

Remark that we do not assume that f belongs to $\mathcal{F}$.

1.3. **Presentation of the results.** In both settings, we give a concentration inequality on the risk of estimators in monodimensionnal models of the form:

$$\{\alpha\theta, \alpha \in \mathbb{R}\}$$

for a given θ.

This results motivates an algorithm that performs iterative feature selection in order to perform regression estimation. We will then remark that the selection procedure gives the guarantee that every selected feature actually improves the current estimator.

In the inductive setting, it means that we estimate $f(.)$ by a function $\hat{f} \in \mathcal{F}$, but the selection procedure can only be performed if the statistician knows the marginal distribution $P_{(X)}$ of X under P.

In the transductive case, the estimation of $Y_{N+1}, ..., Y_{2N}$ can be performed by the procedure without any prior knowledge about the marginal distribution of X under P. We also give in this case some generalizations (like the case where the test sample has a different size).

We then briefly show that the technique used to obtain bounds in models of dimension 1 can also be used in more general models.

In a last section, we come back to the assertion that in our method, "every selected feature actually improves the current estimator" and show how this can be interpreted as an oracle inequality.

2. Main theorem in the inductive case, and application to estimation

Hypothesis. In all this section, we assume that $\mathcal{F}$ and P are such that:

$$\forall \theta \in \Theta, P \exp\left[f_\theta(X) Y\right] < +\infty.$$

2.1. **Notations.** For any random variable T we put:

$$\begin{aligned} V(T) &= P\Big[(T - PT)^2\Big] \\ M^3(T) &= P\Big[(T - PT)^3\Big], \end{aligned}$$

and we define for any $\gamma \geq 0$:

$$P_{\gamma T}(d\omega) = \frac{P\left[\exp\left(\gamma T\right) d\omega\right]}{P\left[\exp\left(\gamma T\right)\right]}.$$

For any random variables T, T' and any $\gamma \geq 0$ we put:

$$V_{\gamma T}(T') = P_{\gamma T}\left[\left(T' - P_{\gamma T}T'\right)^2\right]$$
$$M^3_{\gamma T}(T') = P_{\gamma T}\left[\left(T' - P_{\gamma T}T'\right)^3\right].$$

We give now notations that are specific to the inductive setting.

Definition 2.1. We put:

$$R(\theta) = P\left[\left(Y - f_\theta(X)\right)^2\right]$$
$$r(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - f_\theta(X_i)\right)^2,$$

and in this setting, our objective is $f_{\overline{\theta}}$ where:

$$\overline{\theta} = \arg\min_{\theta\in\Theta} R(\theta).$$

Now, we suppose that we are given a finite family of m vectors:

$$\Theta_0 = \{\theta_1, ..., \theta_m\} \subset \Theta.$$

We are going to use the family Θ_0 to estimate the function f, the estimator will be under the form:

$$\hat{f}(x) = \sum_{k=1}^{m} \alpha_k f_{\theta_k}(x),$$

where every α_k will depend on the observations $Z_1, ..., Z_N$. We can think of Θ_0 as a basis of Θ, but actually there is no other assumption about Θ_0 than finiteness.

Every θ_k defines a monodimensional submodel of $\mathcal{F}$:

$$\{f_{\alpha\theta_k}(.), \alpha \in \mathbb{R}\} = \{\alpha f_{\theta_k}(.), \alpha \in \mathbb{R}\}.$$

In a first step, we are going to work on each of these submodels individually. So let us put, for any $k \in \{1, ..., m\}$:

$$\overline{\alpha}_k = \arg\min_{\alpha\in\mathbb{R}} R(\alpha\theta_k) = \frac{P\left[f_{\theta_k}(X)Y\right]}{P\left[f_{\theta_k}(X)^2\right]}$$
$$\hat{\alpha}_k = \arg\min_{\alpha\in\mathbb{R}} r(\alpha\theta_k) = \frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)Y_i}{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2}$$
$$\mathcal{C}_k = \frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2}{P\left[f_{\theta_k}(X)^2\right]}.$$

2.2. **Main result.** The following theorem gives a control of the excess risk of an estimator in the model $\{f_{\alpha\theta_k}(.), \alpha \in \mathbb{R}\}$ for each k. This estimator is not the usual least square estimator $\hat{\alpha}_k$ but $\mathcal{C}_k\hat{\alpha}_k$.

Theorem 2.1. *Let us put:*

$$W_\theta = f_\theta(X)Y - P\left(f_\theta(X)Y\right).$$

Then we have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $k \in \{1, ..., m\}$:

$$R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) \leq \frac{2\log\frac{2m}{\varepsilon}}{N}\frac{V\left(W_{\theta_k}\right)}{P\left[f_{\theta_k}(X)^2\right]} + \frac{\log^3\frac{2m}{\varepsilon}}{N^{\frac{3}{2}}}C_N(P, m, \varepsilon, \theta_k),$$

where we have:

$$C_N(P,m,\varepsilon,\theta_k) = I_{\theta_k}\left(\sqrt{\frac{2\log\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)^2 \frac{\sqrt{2}}{V\left(W_{\theta_k}\right)^{\frac{5}{2}} P\left[f_{\theta_k}(X)^2\right]} + I_{\theta_k}\left(\sqrt{\frac{2\log\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)^4 \frac{\log^2\frac{2m}{\varepsilon}}{\sqrt{N}V\left(W_{\theta_k}\right)^6 P\left[f_{\theta_k}(X)^2\right]},$$

with:

$$I_\theta(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\beta\gamma W_\theta}\left(W_\theta\right) d\beta.$$

The proof of the theorem is given at the end of this section, let us first show how we can use it in order to build an estimator under the form:

$$\hat{f}(.) = \sum_{k=1}^m \alpha_k f_{\theta_k}(.).$$

Actually, the method we will use requires to be able to compute explicitely the upper bound in this theorem. Remark that, with ε and m fixed:

$$C_N(P,m,\varepsilon,\theta_k) \xrightarrow[N\to+\infty]{} \frac{\sqrt{2}\left[M^3\left(W_{\theta_k}\right)\right]^2}{9V\left(W_{\theta_k}\right)^{\frac{5}{2}} P\left[f_{\theta_k}(X)^2\right]}.$$

and so we can choose to consider only the first order term. Another possible choice is to make stronger assumptions on P and Θ_0 that allow to upper bound explicitely $C_N(P,m,\varepsilon,\theta_k)$. For example, if we assume that Y is bounded by C_Y and that f_{θ_k} is bounded by C'_k then W_{θ_k} is bounded by $C_k = 2C_Y C'_k$ and we have (basically):

$$C_N(P,m,\varepsilon,\theta_k) \leq \frac{64\sqrt{2}C_k^2}{9V\left(W_{\theta_k}\right)^{\frac{5}{2}} P\left[f_{\theta_k}(X)^2\right]} + \frac{4096C_k^4\log^3\frac{2m}{\varepsilon}}{81\sqrt{N}V\left(W_{\theta_k}\right)^6 P\left[f_{\theta_k}(X)^2\right]}.$$

The main problem is actually that the first order term contains the quantity $V\left(W_{\theta_k}\right)$ that is not observable, and we would like to be able to replace this quantity by its natural estimator:

$$\hat{V}_k = \frac{1}{N}\sum_{i=1}^N \left[Y_i f_{\theta_k}(X_i) - \frac{1}{N}\sum_{j=1}^N Y_j f_{\theta_k}(X_j)\right]^2.$$

The following theorem justifies this method.

Theorem 2.2. *If we assume that there is a constant c such that:*

$$\forall k \in \{1,...,m\}, P\left[\exp\left(cW_{\theta_k}^2\right)\right] < \infty,$$

we have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $k \in \{1,...,m\}$:

$$R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) \leq \frac{2\log\frac{4m}{\varepsilon}}{N}\frac{\hat{V}_k}{P\left[f_{\theta_k}(X)^2\right]} + \frac{\log\frac{4m}{\varepsilon}}{N^{\frac{3}{2}}}C'_N(P,m,\varepsilon,\theta_k),$$

where we have:

$$\hat{V}_k = \frac{1}{N}\sum_{i=1}^N \left[Y_i f_{\theta_k}(X_i) - \frac{1}{N}\sum_{j=1}^N Y_j f_{\theta_k}(X_j)\right]^2,$$

and:

$$C'_N(P,m,\varepsilon,\theta_k) = C_N\left(P,m,\frac{\varepsilon}{2},\theta_k\right)\log^2\frac{4m}{\varepsilon} + \frac{2\log^{\frac{1}{2}}\frac{2m}{\varepsilon}}{P\left[f_{\theta_k}(X)^2\right]}\left[\sqrt{2V\left(W^2_{\theta_k}\right)} + \frac{\log\frac{2m}{\varepsilon}}{\sqrt{N}V\left(W^2_{\theta_k}\right)}J_{\theta_k}\left(\sqrt{\frac{2\log\frac{2m}{\varepsilon}}{NV\left(W^2_{\theta_k}\right)}}\right)\right] + \frac{2\log^{\frac{1}{2}}\frac{4m}{\varepsilon}}{P\left[f_{\theta_k}(X)^2\right]}\left[\sqrt{2V\left(W_{\theta_k}\right)} + \frac{\log^2\frac{2m}{\varepsilon}}{\sqrt{N}V\left(W_{\theta_k}\right)^3}I_{\theta_k}\left(\sqrt{\frac{2\log\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)\right]$$
$$\left[\left|\frac{2}{N}\sum_{i=1}^{N}Y_i f_{\theta_k}(X_i)\right|\sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{4m}{\varepsilon}}{N}} + \frac{\log^{\frac{5}{2}}\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3}I_{\theta_k}\left(\sqrt{\frac{2\log\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)\right]$$

and:

$$J_\theta(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\gamma\beta W^2_{\theta_k}}\left(W^2_\theta\right)d\beta.$$

2.3. **Application to regression estimation.**

2.3.1. *Interpretation of theorems 2.1 and 2.2 in terms of confidence intervals.*

Definition 2.2. Let us put, for any $(\theta,\theta')\in\Theta^2$:

$$d_P(\theta,\theta') = \sqrt{P_{(X)}\left[\left(f_\theta(X)-f_{\theta'}(X)\right)^2\right]} = \sqrt{P_{(X)}\left(\langle\theta-\theta',\Psi(X)\rangle^2\right)}.$$

Let also $\|.\|_P$ denote the norm associated with this distance, $\|\theta\|_P = d_P(\theta,0)$, and $\langle.,.\rangle_P$ the associated scalar product:

$$\langle\theta,\theta'\rangle_P = P\left[f_\theta(X)f_{\theta'}(X)\right].$$

Because $\overline{\alpha}_k = \arg\min_{\alpha\in\mathbb{R}} R(\alpha\theta_k)$ we have:

$$R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) = d^2_P(\mathcal{C}_k\hat{\alpha}_k\theta_k,\overline{\alpha}_k\theta_k).$$

So the theorem can be written:

$$P^{\otimes N}\left\{\forall k\in\{1,...,m\},\quad d^2_P(\mathcal{C}_k\hat{\alpha}_k\theta_k,\overline{\alpha}_k\theta_k)\leq\beta(\varepsilon,k)\right\}\geq 1-\varepsilon,$$

where $\beta(\varepsilon,k)$ is the bound given by theorem 2.1 or more likely by theorem 2.2.

Now, note that $\overline{\alpha}_k\theta_k$ is the orthogonal projection of:

$$\overline{\theta} = \arg\min_{\theta\in\Theta} R(\theta)$$

onto the space $\{\alpha\theta_k,\alpha\in\mathbb{R}\}$, with respect to the inner product $\langle.,.\rangle_P$:

$$\overline{\alpha}_k = \arg\min_{\alpha\in\mathbb{R}} d_P\left(\alpha\theta_k,\overline{\theta}\right).$$

Definition 2.3. We define, for any k and ε:

$$\mathcal{CR}(k,\varepsilon) = \left\{\theta\in\Theta : \left|\left\langle\theta-\mathcal{C}_k\hat{\alpha}_k\theta_k,\frac{\theta_k}{\|\theta_k\|_P}\right\rangle_P\right|\leq\sqrt{\beta(\varepsilon,k)}\right\}.$$

Then the theorem is equivalent to the following corollary.

Corollary 2.3. *We have:*

$$P^{\otimes N}\left[\forall k\in\{1,...,m\},\overline{\theta}\in\mathcal{CR}(k,\varepsilon)\right]\geq 1-\varepsilon.$$

In other words: $\bigcap_{k\in\{1,...,m\}}\mathcal{CR}(k,\varepsilon)$ is a confidence region at level ε for $\overline{\theta}$.

Definition 2.4. We write $\Pi_P^{k,\varepsilon}$ the orthogonal projection into $\mathcal{CR}(k,\varepsilon)$ with respect to the distance d_P.

2.3.2. *The algorithm.* The previous formulation of theorem 2.1 motivates the following iterative algorithm:

- choose $\theta(0) \in \Theta$, for example, $\theta(0) = 0$;
- at step $n \in \mathbb{N}^*$, we have: $\theta(0), ..., \theta(n-1)$. Choose $k(n) \in \{1, ..., m\}$ (this choice can of course be data dependent), and take:

$$\theta(n) = \Pi_P^{k(n),\varepsilon} \theta(n-1);$$

- we can use the following stopping rule: $\|\theta(n-1) - \theta(n)\|_P^2 \leq \kappa$ where $0 < \kappa < \frac{1}{N}$.

Definition 2.5. Let n_0 denote the stopping step, and:

$$\hat{f}(.) = f_{\theta(n_0)}(.)$$

the corresponding function.

2.3.3. *Results and comments on the algorithm.*

Theorem 2.4. *We have:*

$$P^{\otimes N}\left[\forall n \in \{1, ..., n_0\}, R\left[\theta\left(n\right)\right] \leq R\left[\theta\left(n-1\right)\right] - d_P^2\left(\theta(n), \theta(n-1)\right)\right] \geq 1 - \varepsilon.$$

Proof. This is just a consequence of the preceding corollary. Let us assume that:

$$\forall k \in \{1, ..., m\}, R(\mathcal{C}_k \hat{\alpha}_k \theta_k) - R(\overline{\alpha}_k \theta_k) \leq \beta(\varepsilon, k)$$

Let us choose $n \in \{1, ..., n_0\}$. We have, for a $k \in \{1, ..., m\}$:

$$\theta(n) = \Pi_P^{k,\varepsilon} \theta(n-1),$$

where $\Pi_P^{k,\varepsilon}$ is the projection into a convex set that contains $\overline{\theta}$. This implies that:

$$\left\langle \theta(n) - \theta(n-1), \overline{\theta} - \theta(n) \right\rangle_P \geq 0,$$

or:

$$d_P^2(\theta(n-1), \overline{\theta}) \geq d_P^2(\theta(n), \overline{\theta}) + d_P^2(\theta(n-1), \theta(n)),$$

that can be written:

$$R\left[\theta(n-1)\right] - R(\overline{\theta}) \geq R\left[\theta(n)\right] - R(\overline{\theta}) + d_P^2(\theta(n-1), \theta(n)).$$

□

Actually, the main point in the motivation of the algorithm is that, with probability at least $1 - \varepsilon$, whatever the current value $\theta(n) \in \Theta$, whatever the feature $k \in \{1, ..., m\}$ (even chosen on the basis of the data), $\Pi_P^{k,\varepsilon} \theta(n)$ is a better estimator than $\theta(n)$.

So we can choose $k(n)$ as we want in the algorithm. For example, theorem 2.4 motivates the choice:

$$k(n) = \arg\max_k d_P^2\left(\theta(n-1), \mathcal{CR}(k, \varepsilon)\right).$$

This version of the algorithm is detailed in figure 1. If looking for the exact maximum of

$$d_P(\theta(n-1), \mathcal{CR}(k, \varepsilon))$$

with respect to k is too computationnaly intensive we can use any heuristic to choose $k(n)$, or even skip this maximization and take:

$$k(1) = 1, ..., k(m) = m, k(m+1) = 1, ..., k(2m) = m, ...$$

FIGURE 1. Detailed version of the feature selection algorithm.

We have $\varepsilon > 0$, $\kappa > 0$, N observations $(X_1, Y_1), ..., (X_N, Y_N)$, m features $f_{\theta_1}(.), ..., f_{\theta_m}(.)$ and $\theta(0) = (\theta_1(0), ..., \theta_m(0)) = (0, ..., 0) \in \mathbb{R}^m$. Compute at first every $\hat{\alpha}_k$ and $\beta(\varepsilon, k)$ for $k \in \{1, ..., m\}$. Set $n \leftarrow 0$.

Repeat:

- set $n \leftarrow n + 1$;
- set $best\_improvement \leftarrow 0$ and $\theta(n) \leftarrow \theta(n-1)$;
- for $k \in \{1, ..., m\}$, compute:

$$v_k = P\left[f_{\theta_k}(X)^2\right],$$

$$\gamma_k \leftarrow \hat{\alpha}_k - \frac{1}{v_k} \sum_{j=1}^{m} \theta_j(n) P\left[f_{\theta_j}(X) f_{\theta_k}(X)\right],$$

$$\delta_k \leftarrow v_k \Big(|\gamma_k| - \beta(\varepsilon, k) \Big)_+^2,$$

  and if $\delta_k > best\_improvement$, set:

$$best\_improvement \leftarrow \delta_k,$$

$$k(n) \leftarrow k;$$

- if $best\_improvement > 0$:

$$\theta_{k(n)}(n) \leftarrow \theta_{k(n)}(n) + sgn(\gamma_k) \Big(|\gamma_k| - \beta(\varepsilon, k) \Big)_+;$$

until $best\_improvement < \kappa$ (where $sgn(x) = -1$ if $x \leq 0$ and 1 otherwise).

Return the estimator:

$$\hat{f}(.) = \sum_{k=1}^{m} \theta_k(n) f_{\theta_k}(.).$$

Such a procedure could look similar to the famous Widrow-Hoff algorithm [15] (also known as ADALINE), which estimates the function $f(.)$ by an estimator under the form:

$$\sum_{k=1}^{m} \alpha_k f_{\theta_k}(.),$$

and updates the α_k sequentially by a gradient descent strategy. Actually, there are two major differences: first, the gradient descent requires the calibration of a parameter $\eta > 0$, that is avoided here, then, ADALINE is only a way to compute the usual least square estimator, and has absolutely no guarantees against overlearning if the family $\{f_{\theta_1}, ..., f_{\theta_m}\}$ is too large.

Example 2.1. Let us assume that $\mathcal{X} = [0, 1]$ and let us put $\Theta = \mathbb{L}_2(P_{(X)})$. Let $(\theta_k)_{k \in \mathbb{N}^*}$ be an orthonormal basis of Θ and we simply take, for any x and θ:

$$f_\theta(x) = \theta(x).$$

The choice of m should not be a problem, the algorithm avoiding itself overlearning we can take a large value of m like $m = N$ (see later for more details). In this setting, the algorithm is a procedure for (soft) thresholding of coefficients. In the particular case of a wavelets basis, see Kerkyacharian and Picard [9] or Donoho and Johnstone [8] for a presentation of wavelets coefficient thresholding. Here, the threshold is not necessarily the same for every coefficient. We can remark that the

sequential projection on every k is sufficient here:

$$k(1) = 1, ..., k(m) = m,$$

after that $\theta(m+n) = \theta(m)$ for every $n \in \mathbb{N}$ (because all the directions of the different projections are orthogonals).

Actually, in the case given in the example, it is possible to prove that the estimator is able to adapt itself to the regularity of the function to achieve a good mean rate of convergence. More precisely, if we assume that the true regression function has an (unknown) regularity β, then it is possible to choose m and ε in such a way that the rate of convergence is:

$$N^{\frac{-2\beta}{2\beta+1}} \log N.$$

We prove this point in the last section of this paper.

2.4. **An extension to the case of Support Vector Machines.** Thanks to a method due to Seeger [14], it is possible to extend this method to the case where the set Θ_0 is data dependent in the following way:

$$\Theta_0(Z_1, ..., Z_N, N) = \bigcup_{i=1}^{N} \Theta_0(Z_i, N),$$

where for any $z \in \mathcal{X} \times \mathbb{R}$, the cardinal of the set $\Theta_0(z, N)$ depends only on N, not on z. We will write $m'(N)$ this cardinal. So we have:

$$|\Theta_0(Z_1, ..., Z_N, N)| \leq N \left|\Theta_0(Z_i, N)\right| = N m'(N).$$

We put:

$$\Theta_0(Z_i, N) = \left\{\theta_{i,1}, ..., \theta_{i,m'(N)}\right\}.$$

In this case, we need some adaptations of our previous notations. We put, for $i \in \{1, ..., N\}$:

$$r_i(\theta) = \frac{1}{N-1} \sum_{\substack{j \in \{1, ..., N\} \\ j \neq i}} \left(Y_j - f_\theta(X_j)\right)^2.$$

For any $(i,k) \in \{1, ..., N\} \times \{1, ..., m'(N)\}$, we write:

$$\hat{\alpha}_{i,k} = \arg\min_{\alpha \in \mathbb{R}} r_i(\alpha \theta_{i,k}) = \frac{\sum_{j \neq i} f_{\theta_{i,k}}(X_j) Y_j}{\sum_{j \neq i} f_{\theta_{i,k}}(X_j)^2}$$

$$\overline{\alpha}_{i,k} = \arg\min_{\alpha \in \mathbb{R}} R(\alpha \theta_{i,k}) = \frac{P\left[f_{\theta_{i,k}}(X) Y\right]}{P\left[f_{\theta_{i,k}}(X)^2\right]}$$

$$\mathcal{C}_{i,k} = \frac{\frac{1}{N-1} \sum_{j \neq i} f_{\theta_{i,k}}(X_j)^2}{P\left[f_{\theta_{i,k}}(X)^2\right]}.$$

Theorem 2.5. *We have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $k \in \{1, ..., m'(N)\}$ and $i \in \{1, ..., N\}$:*

$$R(\mathcal{C}_{i,k} \hat{\alpha}_{i,k} \theta_{i,k}) - R(\overline{\alpha}_{i,k} \theta_{i,k}) \leq \frac{2 \log \frac{2Nm'(N)}{\varepsilon}}{N-1} \frac{V\left(W_{\theta_{i,k}}\right)}{P\left[f_{\theta_{i,k}}(X)^2\right]} + \frac{\log^3 \frac{2Nm'(N)}{\varepsilon}}{(N-1)^{\frac{3}{2}}} C_{N-1}(P, Nm'(N), \varepsilon, \theta_{i,k}).$$

We can use this theorem to build an estimator using the algorithm described in the previous subsection, with obvious changes in the notations.

Example 2.2. Let us consider the case where Θ is a Hilbert space with scalar product $\langle .,. \rangle$, and:

$$f_\theta(x) = \langle \theta, \Psi(x) \rangle$$

for any $\theta \in \Theta$ and $x \in \mathcal{X}$, where Ψ is an application $\mathcal{X} \to \Theta$. Let us put $\Theta_0[(x,y),N] = \{\Psi(x)\}$. In this case we have $m'(N) = 1$ and:

$$\hat{f}(.) = \sum_{i=1}^{N} \alpha_{i,1} \langle \Psi(X_i), \Psi(.) \rangle .$$

Let us define,

$$K(x,x') = \langle \Psi(x), \Psi(x') \rangle ,$$

the function K is called the kernel, and:

$$I = \{1 \leq i \leq N : \alpha_{i,1} \neq 0\},$$

that is called the set of support vectors. Then the estimate has the form of a support vector machine (SVM):

$$\hat{f}(.) = \sum_{i \in I} \alpha_{i,1} K(X_i, .).$$

SVM where first introduced by Boser, Guyon and Vapnik [2] in the context of classification, and then generalized by Vapnik [12] to the context of regression estimation. For a general introduction to SVM, see also Cristianini and Shawe-Taylor [7] and Catoni [5].

Example 2.3. A widely used kernel is the gaussian kernel:

$$K_\gamma(x,x') = \exp\left(-\gamma \frac{d^2(x,x')}{2}\right),$$

where $d(.,.)$ is some distance over the space $\mathcal{X}$ and $\gamma > 0$. But in practice, the choice of the parameter γ is difficult. A way to solve this problem is to introduce multiscale SVM. We simply take Θ as the set of all bounded functions $\mathcal{X} \to \mathbb{R}$, and for any x and θ:

$$f_\theta(x) = \theta(x).$$

Now, let us put:

$$\Theta_0[(x,y),N] = \{K_2(x,.), K_{2^2}(X,.), ..., K_{2^{m'(N)}}(x,.)\} .$$

In this case, we obtain an estimator of the form:

$$\hat{f}(.) = \sum_{k=1}^{m'(N)} \sum_{i \in I_k} \alpha_{i,k} K_{2^k}(X_i, .),$$

that could be called multiscale SVM. Remark that we can use this technique to define SVM using simultaneously different kernels (not necessarily the same kernel at different scales). For example, in order to imitate the oscillation of wavelets, we can introduce a more sophisticated SVM estimator, based on the kernel family:

$$K_{\gamma,\gamma'}(x,x') = \exp\left(-2^{2\gamma}(x-x')^2\right) \cos\left(2^{\gamma+\gamma'-1}(x-x')\right)$$

for $\gamma \in \{1, ..., m_1\}$, $\gamma' \in \{1, ..., m_2\}$ (note that $m'(N) = m_1 m_2$).

2.5. **Proof of the theorems.** In a first time, we prove a lemma that is the basis of proofs of theorems 2.1 and 2.5.

Lemma 2.6. *We have, for any $\theta \in \Theta$, $\gamma > 0$ and $\eta \geq 0$:*

$$P\exp\left(\gamma W_\theta - \eta\right) = \exp\left\{\frac{\gamma^2}{2}V\left(W_\theta\right) + \frac{\gamma^3}{2}\int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta}\left(W_\theta\right)d\beta - \eta\right\},$$

and

$$P\exp\left(-\gamma W_\theta - \eta\right) = \exp\left\{\frac{\gamma^2}{2}V\left(W_\theta\right) - \frac{\gamma^3}{2}\int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta}\left(W_\theta\right)d\beta - \eta\right\}.$$

Proof. For the first equality, we write:

$$\begin{aligned}
\log P\exp\left(\gamma W_\theta - \eta\right) &= \log P\exp\left(\gamma W_\theta\right) - \eta \\
&= \int_0^\gamma P_{\beta W_\theta}\left(W_\theta\right)d\beta - \eta = \int_0^\gamma (\gamma-\beta)V_{\beta W_\theta}\left(W_\theta\right)d\beta - \eta \\
&= \frac{\gamma^2}{2}V\left(W_\theta\right) + \int_0^\gamma \frac{(\gamma-\beta)^2}{2}M^3_{\beta W_\theta}\left(W_\theta\right)d\beta - \eta \\
&= \frac{\gamma^2}{2}V\left(W_\theta\right) + \frac{\gamma^3}{2}\int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta}\left(W_\theta\right)d\beta - \eta.
\end{aligned}$$

For the reverse equality, the proof is exactly the same, replacing γ by $-\gamma$. □

We can now give the proof of both theorems.

Proof of theorem 2.1. Let us choose $k \in \{1, ..., m\}$, for any $\lambda_k > 0$ and $\eta_k \geq 0$ we have:

$$\begin{aligned}
P^{\otimes N}\exp&\left\{\frac{\lambda_k}{N}\sum_{i=1}^N \left[Y_i f_{\theta_k}(X_i) - P\left(Y f_{\theta_k}(X)\right)\right] - \eta_k\right\} \\
&= \left\{P\exp\left[\frac{\lambda_k}{N}W_{\theta_k} - \frac{\eta_k}{N}\right]\right\}^N \\
&= \exp\left[\frac{\lambda_k^2}{2N}V\left(W_{\theta_k}\right) + \frac{\lambda_k^3}{2N^2}\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N}W_{\theta_k}}\left(W_{\theta_k}\right)d\beta - \eta_k\right]
\end{aligned}$$

by the first equality of lemma 2.6. By the same way, using the reverse inequality we obtain:

$$\begin{aligned}
P^{\otimes N}\exp&\left\{\frac{\lambda_k}{N}\sum_{i=1}^N \left[P\left(Y f_{\theta_k}(X)\right) - Y_i f_{\theta_k}(X_i)\right] - \eta_k\right\} \\
&= \exp\left[\frac{\lambda_k^2}{2N}V\left(W_{\theta_k}\right) - \frac{\lambda_k^3}{2N^2}\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N}W_{\theta_k}}\left(W_{\theta_k}\right)d\beta - \eta_k\right].
\end{aligned}$$

So we obtain, for any $k \in \{1, ..., m\}$, for any $\lambda_k > 0$ and $\eta_k \geq 0$:

$$\begin{aligned}
P^{\otimes N}\exp&\left\{\lambda_k\left|\frac{1}{N}\sum_{i=1}^N Y_i f_{\theta_k}(X_i) - P\left(Y f_{\theta_k}(X)\right)\right| - \eta_k\right\} \\
&\leq 2\exp\left[\frac{\lambda_k^2}{2N}V\left(W_{\theta_k}\right) - \eta_k\right]\cosh\left[\frac{\lambda_k^3}{2N^2}\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N}W_{\theta_k}}\left(W_{\theta_k}\right)d\beta\right] \\
&\leq 2\exp\left[\frac{\lambda_k^2}{2N}V\left(W_{\theta_k}\right) - \eta_k + \frac{\lambda_k^6}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N}W_{\theta_k}}\left(W_{\theta_k}\right)d\beta\right)^2\right],
\end{aligned}$$

since, for any $x \in \mathbb{R}$, we have:

$$\cosh(x) \leq \exp\left(\frac{x^2}{2}\right).$$

Now, let us choose $\varepsilon > 0$ and put:

$$\eta_k = \frac{\lambda_k^2}{2N} V\left(W_{\theta_k}\right) + \frac{\lambda_k^6}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N} W_{\theta_k}}\left(W_{\theta_k}\right) d\beta\right)^2 - \log\frac{\varepsilon}{2m}.$$

We obtain:

$$P^{\otimes N}\sum_{k=1}^m \exp\Bigg\{\lambda_k\left|\frac{1}{N}\sum_{i=1}^N Y_i f_{\theta_k}(X_i) - P\left(Y f_{\theta_k}(X)\right)\right| \\ - \frac{\lambda_k^2}{2N} V\left(W_{\theta_k}\right) + \frac{\lambda_k^6}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N} W_{\theta_k}}\left(W_{\theta_k}\right) d\beta\right)^2 + \log\frac{\varepsilon}{2m}\Bigg\} \leq \varepsilon$$

and so:

$$P^{\otimes N}\Bigg[\forall k \in \{1,...,m\}, \quad \left|\frac{1}{N}\sum_{i=1}^N Y_i f_{\theta_k}(X_i) - P\left(Y f_{\theta_k}(X)\right)\right| \\ \leq \frac{\lambda_k}{2N} V\left(W_{\theta_k}\right) + \frac{\lambda_k^5}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N} W_{\theta_k}}\left(W_{\theta_k}\right) d\beta\right)^2 + \frac{\log\frac{2m}{\varepsilon}}{\lambda_k}\Bigg] \geq 1-\varepsilon.$$

Now, we put:

$$\lambda_k = \sqrt{\frac{2N\log\frac{2m}{\varepsilon}}{V\left(W_{\theta_k}\right)}}.$$

We obtain, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $k \in \{1,...,m\}$:

$$\left|\frac{1}{N}\sum_{i=1}^N Y_i f_{\theta_k}(X_i) - P\left(Y f_{\theta_k}(X)\right)\right| \\ \leq \sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{2m}{\varepsilon}}{N}} + \frac{\log^{\frac{5}{2}}\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3}\left(\int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_k}{N} W_{\theta_k}}\left(W_{\theta_k}\right) d\beta\right)^2.$$

For short, we take the notation of the theorem:

$$I_{\theta_k}(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\beta\gamma W_{\theta_k}}\left(W_{\theta_k}\right).$$

Now, dividing both sides by:

$$P\left[f_{\theta_k}(X)^2\right]$$

we obtain:

$$|\hat{\alpha}_k\mathcal{C}_k - \overline{\alpha}_k| \leq \frac{1}{P\left[f_{\theta_k}(X)^2\right]}\left[\sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{2m}{\varepsilon}}{N}} + \frac{I^2_{\theta_k}\left(\frac{\lambda_k}{N}\right)\log^{\frac{5}{2}}\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3}\right].$$

In order to conclude, just remark that:

$$R(\hat{\alpha}_k\mathcal{C}_k\theta_k) - R(\overline{\alpha}_k\theta_k) = |\hat{\alpha}_k\mathcal{C}_k - \overline{\alpha}_k|^2 P\left[f_{\theta_k}(X)^2\right].$$

□

Proof of theorem 2.2. Remark that, for any $\theta \in \Theta$:

$$V\left(W_\theta\right) = P\left(W_\theta^2\right) - P\left(W_\theta\right)^2,$$

we will deal with each term separately. For the first term, let us remark that we obtain the following result that is obtained exactly as lemma 2.6. For any $\theta \in \Theta$:

$$P \exp\left\{\gamma\left[P\left(W_\theta^2\right) - W_\theta^2\right] - \eta\right\} = \exp\left\{\frac{\gamma^2}{2} V\left(W_\theta^2\right) + \frac{\gamma^3}{2}\int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta^2}\left(W_\theta^2\right) d\beta - \eta\right\}.$$

Let us apply this result to every θ_k for $k \in \{1, ..., m\}$:

$$P^{\otimes N} \exp\left\{\lambda_k\left[P\left(W_{\theta_k}^2\right) - \frac{1}{N}\sum_{i=1}^N Y_i^2 f_{\theta_k}(X_i)^2\right] - \eta_k\right\} = \exp\left\{\frac{\lambda_k^2}{2N} V\left(W_{\theta_k}^2\right) + \frac{\lambda_k^3}{2N} J_k\left(\frac{\lambda_k}{N}\right) - \eta_k\right\},$$

where:

$$J_\theta(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_{\theta_k}^2}\left(W_\theta^2\right) d\beta.$$

Taking:

$$\eta_k = \frac{\lambda_k^2}{2N} V\left(W_{\theta_k}^2\right) + \frac{\lambda_k^3}{2N^2} J_{\theta_k}\left(\frac{\lambda_k}{N}\right) + \log\frac{2m}{\varepsilon}$$

and:

$$\lambda_k = \sqrt{\frac{2N\log\frac{2m}{\varepsilon}}{V\left(W_{\theta_k}^2\right)}}$$

we obtain that the following inequality is satisfied with $P^{\otimes N}$-probability at least $1 - \frac{\varepsilon}{2}$, for any k:

$$P\left(W_{\theta_k}^2\right) \leq \frac{1}{N}\sum_{i=1}^N Y_i^2 f_{\theta_k}(X_i)^2 + \sqrt{\frac{2V\left(W_{\theta_k}^2\right)\log\frac{2m}{\varepsilon}}{N}} + \frac{\log\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}^2\right)} J_{\theta_k}\left(\sqrt{\frac{2\log\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}^2\right)}}\right) = \frac{1}{N}\sum_{i=1}^N Y_i^2 f_{\theta_k}(X_i)^2 + \mathcal{A}_k \tag{2.1}$$

for short. Now, we try to upper bound the second term, $-P\left(W_\theta\right)^2$. Remark that, for any θ:

$$\left(\frac{1}{N}\sum_{i=1}^N Y_i f_\theta(X_i)\right)^2 - P\left(W_\theta\right)^2 = \left(\frac{1}{N}\sum_{i=1}^N Y_i f_\theta(X_i) - P\left(W_\theta\right)\right)\left(\frac{1}{N}\sum_{i=1}^N Y_i f_\theta(X_i) + P\left(W_\theta\right)\right) \leq \left|\frac{1}{N}\sum_{i=1}^N Y_i f_\theta(X_i) - P\left(W_\theta\right)\right| \left\{2\left|\frac{1}{N}\sum_{i=1}^N Y_i f_\theta(X_i)\right| + \left|\frac{1}{N}\sum_{i=1}^N Y_i f_\theta(X_i) - P\left(W_\theta\right)\right|\right\}.$$

Remember that in the proof of theorem 2.1 we got the upper bound, with probability at least $1-\frac{\varepsilon}{2}$, for any k:

$$\left|\frac{1}{N}\sum_{i=1}^{N}Y_i f_{\theta_k}(X_i)-P\left(Yf_{\theta_k}(X)\right)\right| \leq \sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{4m}{\varepsilon}}{N}}+\frac{\log^{\frac{5}{2}}\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3}I_{\theta_k}\left(\sqrt{\frac{2\log\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)^2,$$

that gives:

$$\begin{aligned}(2.2)\quad -P\left(W_{\theta_k}\right)^2 \leq & -\left(\frac{1}{N}\sum_{i=1}^{N}Y_i f_{\theta_k}(X_i)\right)^2 \\ & +\left\{\sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{4m}{\varepsilon}}{N}}+\frac{\log^{\frac{5}{2}}\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3}I_{\theta_k}\left(\sqrt{\frac{2\log\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)^2\right\} \\ & \left\{2\left|\frac{1}{N}\sum_{i=1}^{N}Y_i f_{\theta_k}(X_i)\right|+\right. \\ & \left.\sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{4m}{\varepsilon}}{N}}+\frac{\log^{\frac{5}{2}}\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3}I_{\theta_k}\left(\sqrt{\frac{2\log\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)^2\right\} \\ & =-\left(\frac{1}{N}\sum_{i=1}^{N}Y_i f_{\theta}(X_i)\right)^2+\mathcal{B}_k.\end{aligned}$$

for short. Let us combine inequalities 2.1 and 2.2. We obtain that, with probability at least $1-\varepsilon$, for every k we have:

$$\begin{aligned}V\left(W_{\theta_k}\right) &= P\left(W_{\theta_k}^2\right)-P\left(W_{\theta_k}\right)^2 \\ &\leq \frac{1}{N}\sum_{i=1}^{N}Y_i^2 f_{\theta_k}(X_i)^2-\left(\frac{1}{N}\sum_{i=1}^{N}Y_i f_{\theta_k}(X_i)\right)^2+\mathcal{A}_k+\mathcal{B}_k \\ &= \hat{V}_k+\mathcal{A}_k+\mathcal{B}_k.\end{aligned}$$

□

Proof of theorem 2.5. This proof is a variant of the proof of theorem 2.1, the method it uses is due to Seeger [14]. Let us define, for any $i\in\{1,...,N\}$:

$$P_i(.)=P^{\otimes N}\left(.|Z_i\right).$$

Let us choose $(i,k)\in\{1,...,N\}\times\{1,...,m'(N)\}$, for any $\lambda_{i,k}=\lambda_{i,k}(Z_i)>0$ and $\eta_{i,k}=\eta_{i,k}(Z_i)\geq 0$ we have:

$$\begin{aligned}P_i\exp&\left\{\frac{\lambda_{i,k}}{N-1}\sum_{j\neq i}\left[Y_j f_{\theta_{i,k}}(X_j)-P\left(Yf_{\theta_{i,k}}(X)\right)\right]-\eta_{i,k}\right\} \\ &\leq \exp\left[\frac{\lambda_{i,k}}{2(N-1)}V\left(W_{\theta_{i,k}}\right)\right. \\ &\left.+\frac{\lambda_{i,k}^3}{2(N-1)^2}\int_0^1(1-\beta)^2 M^3_{\frac{\beta\lambda_{i,k}}{N-1}W_{\theta_{i,k}}}\left(W_{\theta_{i,k}}\right)d\beta-\eta_{i,k}\right]\end{aligned}$$

by the first equality of lemma 2.6. In the same way, we obtain the reverse inequality and, combining both results, for any $(i,k) \in \{1,...,N\} \times \{1,...,m'(N)\}$, for any $\lambda_{i,k} > 0$ and $\eta_{i,k} \geq 0$:

$$P_i \exp\left\{ \lambda_{i,k} \left| \frac{1}{N-1} \sum_{j \neq i} Y_j f_{\theta_{i,k}}(X_j) - P\left(Y f_{\theta_{i,k}}(X)\right) \right| - \eta_{i,k} \right\}$$
$$\leq 2 \exp\left[\frac{\lambda_{i,k}^2}{2(N-1)} V\left(W_{\theta_{i,k}}\right) - \eta_{i,k} \right] \cosh\left[\frac{\lambda_{i,k}^3}{2(N-1)^2} I_{i,k} \right]$$
$$\leq 2 \exp\left[\frac{\lambda_{i,k}^2}{2(N-1)} V\left(W_{\theta_{i,k}}\right) - \eta_{i,k} + \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2 \right],$$

where:

$$I_{i,k} = \int_0^1 (1-\beta)^2 M^3_{\frac{\beta\lambda_{i,k}}{N} W_{\theta_{i,k}}}\left(W_{\theta_{i,k}}\right) d\beta$$

for short. Now, let us choose $\varepsilon > 0$ and put:

$$\eta_{i,k} = \frac{\lambda_{i,k}^2}{2(N-1)} V\left(W_{\theta_{i,k}}\right) + \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2 - \log \frac{\varepsilon}{2Nm'(N)}.$$

We obtain:

$$P^{\otimes N} \sum_{i=1}^{N} \sum_{k'=1}^{m'(N)} \exp\Bigg\{ \lambda_{i,k} \left| \frac{1}{N-1} \sum_{j \neq i} Y_j f_{\theta_{i,k}}(X_j) - P\left(Y f_{\theta_{i,k}}(X)\right) \right|$$
$$- \frac{\lambda_{i,k}^2}{2(N-1)} V\left(W_{\theta_{i,k}}\right) - \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2 + \log \frac{\varepsilon}{2Nm'(N)} \Bigg\}$$
$$= P^{\otimes N} \sum_{i=1}^{N} \sum_{k'=1}^{m'(N)} P_i \exp\Bigg\{ \lambda_{i,k} \left| \frac{1}{N-1} \sum_{j \neq i} Y_j f_{\theta_{i,k}}(X_j) - P\left(Y f_{\theta_{i,k}}(X)\right) \right|$$
$$- \frac{\lambda_{i,k}^2}{2(N-1)} V\left(W_{\theta_{i,k}}\right) - \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2 + \log \frac{\varepsilon}{2Nm'(N)} \Bigg\} \leq \varepsilon.$$

Now, we put:

$$\lambda_{i,k} = \sqrt{\frac{2N \log \frac{2Nm'(N)}{\varepsilon}}{V\left(W_{\theta_{i,k}}\right)}},$$

and achieve the proof exactly as for theorem 2.1. □

3. Simulations in the inductive case

3.1. **Description of the example.** Here, we assume that we have:

$$Y_i = f(X_i) + \xi_i$$

for $i \in \{1,...,N\}$ with $N = 2^{10} = 1024$, where the variables $X_i \in [0,1] \subset \mathbb{R}$ are i.i.d. from a uniform distribution $\mathcal{U}(0,1)$ (and we assume that the statistician knows this point), the η_i are i.i.d. from a gaussian distribution $\mathcal{N}(0,\sigma)$ and independant from the X_i. The statistician observes $(X_1,Y_1),...,(X_N,Y_N)$ and wants to estimate the regression function f.

We will use three estimations methods. The first one will be an SVM obtained by the algorithm described previously, the second one a thresholded wavelets estimate also obtained by this algorithm, and we will compare both estimators to a "classical" thresholded wavelet estimate, as given by Kerkyacharian and Picard [9].

3.2. **The estimators.**

3.2.1. *Thresholded wavelets estimators.* Let us describe briefly the thresholded wavelet estimator. Let (φ, ψ) be the father wavelet and the mother wavelet, and:

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x + k)$$

for $k \in \{0, ..., 2^j - 1\} = S_j$. For the sake of simplicity, let us write:

$$\psi_{-1,k}(x) = \varphi(x)$$

for $k \in \{0\} = S_{-1}$.

Here, we will use the Haar basis, with:

$$\varphi(x) = \mathbb{1}_{[0,1]}(x)$$
$$\psi(x) = \mathbb{1}_{\left[0,\frac{1}{2}\right]}(x) - \mathbb{1}_{\left[\frac{1}{2},1\right]}(x).$$

In the general case, we should use warped wavelets (for more details, see Kerkyacharian and Picard [9]): we put $F(x) = P(X \leq x)$, and:

$$\hat{\beta}_{j,k} = \frac{1}{N}\sum_{i=1}^{N} Y_i \psi_{j,k}(F(X_i)).$$

Just remark that the use of this method implies some assumptions about F that are not required by our algorithm (here again, see Kerkyacharian and Picard [9]).

In the case of the example, we will have:

$$\hat{\beta}_{j,k} = \frac{1}{N}\sum_{j=1}^{N} Y_i \psi_{j,k}(X_i).$$

For a given $\kappa \geq 0$ and $J \in \mathbb{N}$, we take:

$$\tilde{f}_J(.) = \sum_{j=-1}^{J}\sum_{k\in S_j} \hat{\beta}_{j,k}\mathbb{1}(|\hat{\beta}_{j,k}| \geq \kappa t_N)\psi_{j,k}(.)$$

where:

$$t_N = \sqrt{\frac{\log N}{N}}.$$

Actually, we must choose J in such a way that:

$$2^J \sim t_N^{-1}.$$

When $\kappa = 0$ we obtain a classical wavelet estimator, and when $\kappa > 0$ we obtain a thresholded wavelet estimator, this is what we are going to do.

Here, we choose $\kappa = 0.5$ and $J = 7$.

3.2.2. *Wavelet estimators with our algorithm.* Here, we use the same family of functions, and we apply the algorithm given in subsection 2.3. So we take:

$$m = 2^J = 128.$$

We change only one thing in the method in order to obtain faster computations: here, applying the central limit theorem, we replace the theoretical confidance interval by its asymptotic gaussian approximation.

More precisely:

$$\sqrt{N}\frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)Y_i - P\left[f_{\theta_k}(X)Y\right]}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(f_{\theta_k}(X_i)Y_i - \frac{1}{N}\sum_{j=1}^{N} f_{\theta_k}(X_j)Y_j\right)^2}} \rightsquigarrow \mathcal{N}(0,1).$$

FIGURE 2. Values of t_i and c_i in the fonction $Blocks(.)$.

i	1	2	3	4	5	6	7	8	9	10	11
c_i	4	-5	3	-4	5	4.2	-2.1	4.3	-3.1	2.1	-4.2
t_i	0.10	0.13	0.15	0.23	0.25	0.40	0.44	0.65	0.76	0.78	0.81

We put:

$$v_{k,N} = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(f_{\theta_k}(X_i)Y_i - \frac{1}{N}\sum_{j=1}^{N} f_{\theta_k}(X_j)Y_j\right)^2}}{\sqrt{N}}.$$

We obtain:

$$\left(\mathcal{C}_k\hat{\alpha}_k - \overline{\alpha}_k\right)\frac{P\left[f_{\theta_k}(X)^2\right]}{v_{k,N}} \rightsquigarrow \mathcal{N}(0,1),$$

or:

$$\left(R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k)\right)\frac{P\left[f_{\theta_k}(X)^2\right]}{v_{k,N}^2} \rightsquigarrow \chi_1^2,$$

and so we use the confidence interval for $\overline{\alpha}_k$:

$$\left[\mathcal{C}_k\hat{\alpha}_k \pm \frac{v_{k,N}}{P\left[f_{\theta_k}(X)^2\right]}q_{1-\frac{\varepsilon}{2m'(N)}}\right]$$

where q_α is the α-quantile of $\mathcal{N}(0,1)$.

Remark that the numerical results are not very different if we use the confidence interval given by theorem 2.1.

Moreover, let us remark that the union bound are always "pessimistic", and that we use a union bound argument over all the m models despite only a few of them are effectively used in the estimator. So, we propose to actually use the individual confidence interval for each model:

$$\left[\mathcal{C}_k\hat{\alpha}_k \pm \frac{v_{k,N}}{P\left[f_{\theta_k}(X)^2\right]}q_{1-\frac{\varepsilon}{2}}\right]$$

instead of the theoretical union bound interval.

3.2.3. *SVM estimator.* Here, we use the multiscale SVM estimator described in example 2.3 of subsection 2.4, with kernel:

$$K_\gamma(x,x') = \exp\left(-(2^\gamma x - 2^\gamma x')^2\right) = \exp\left(-2^{2\gamma}(x-x')^2\right)$$

and $\gamma \in \{1,...,m'(N)\}$ where $m'(N) = 6$.

We use the same gaussian approximation than in the previous example, and the individuals confidence intervals.

3.3. **Experiments and results.** The simulations were realized with the R software [11].

For the experiments, we use the following functions f that are some of the functions used by Donoho and Johnstone for experiments on wavelets, for example in [8], and by a lot of authors since then:

$$Doppler(t) = u\sqrt{t(1-t)}\sin\frac{2\pi(1+v)}{t+v} \quad \text{where } u = 2 \text{ and } v = 0.05$$

$$HeaviSine(t) = \frac{1}{4}\Big[4\sin 4\pi t - sgn(t-0.3) - sgn(0.72-t)\Big]$$

$$Blocks(t) = \frac{1}{4}\sum_{i=1}^{11} c_i \mathbb{1}_{(t_i,+\infty)}(t)$$

where $sgn(t)$ is the sign of t (say -1 if $t \leq 0$ and $+1$ otherwise). The values of the c_i and t_i are given in figure 2.

FIGURE 3. Results of the experiments. For each experiment, we give the mean risk (R) and the mean excess risk ($R-\sigma^2$) for each estimator.

Function $f(.)$	s.d. σ	standard thresholded wavelets	thresh. wav. with our method	multiscale SVM
Doppler	0.3	0.158 / 0.068	0.151 / 0.061	0.149 / 0.059
HeaviSine	0.3	0.154 / 0.064	0.138 / 0.048	0.129 / 0.039
Blocks	0.3	0.150 / 0.060	0.146 / 0.056	0.159 / 0.069
Doppler	1	1.142 / 0.142	1.114 / 0.114	1.091 / 0.091
HeaviSine	1	1.156 / 0.156	1.084 / 0.084	1.055 / 0.055
Blocks	1	1.155 / 0.155	1.105 / 0.105	1.104 / 0.104

We consider 6 experiments (for the three regression functions and two different values for σ, 0.3 and 1). We choose ε=10%. We repeat each experiment 20 times. We give the results in figure 3.

The result of thresholding wavelets following [9] or using our algorithm is comparable. However, our thresholding method gives best results, especially when the noise level is significant. The main advantage of our method is that is is self-contained: in the "standard" thresholding, we have to choose the parameter κ. Here, the choice $\kappa = 0.5$ seemed to give the better results, but this choice was possible only because we knew the regression function in these simulations. In real life problems, the choice of κ could be more problematic.

SVM gave best results, except in the case where $f = Blocks$. But the main advantage of SVM is that it is much easier to generalize in the case where $\mathcal{X}$ is not $\mathbb{R}$ or an interval of $\mathbb{R}$, but for example in the case where $\mathcal{X} = \mathbb{R}^n$ with $n \geq 2$. More genarally, let us assume that $\mathcal{X}$ is a metric space for some distance d. We can use SVM with the gaussian kernel ($(x, x') \in \mathcal{X}^2$):

$$K_\gamma(x, x') = \exp\left(-2^{2\gamma}\frac{d^2(x, x')}{2}\right).$$

4. The transductive case

Remark that in this section, we make no longer assumptions about the existence of an exponential moment for $f_\theta(X)Y$.

4.1. **Notations.** Let us recall that we assume that P_{2N} is some exchangeable probability measure on the space $\left((\mathcal{X}\times\mathbb{R})^{2N}, (\mathcal{B}\times\mathcal{B}_\mathbb{R})^{\otimes 2N}\right)$. Let $(X_i, Y_i)_{i=1\dots 2N} = (Z_i)_{i=1\dots 2N}$ denote a random vector distributed according to P_{2N}.

Let us remark that under this condition, the marginal distribution of every Z_i is the same, we will call P this distribution. In the particular case where the observations are i.i.d., we will have $P_{2N} = P^{\otimes 2N}$, but what follows still holds for general exchangeable distributions P_{2N}.

We assume that we observe $(X_i, Y_i)_{i=1\dots N}$ and $(X_i)_{i=N+1\dots 2N}$. In this case, we only focus on the estimation of the values $(Y_i)_{i=N+1\dots 2N}$.

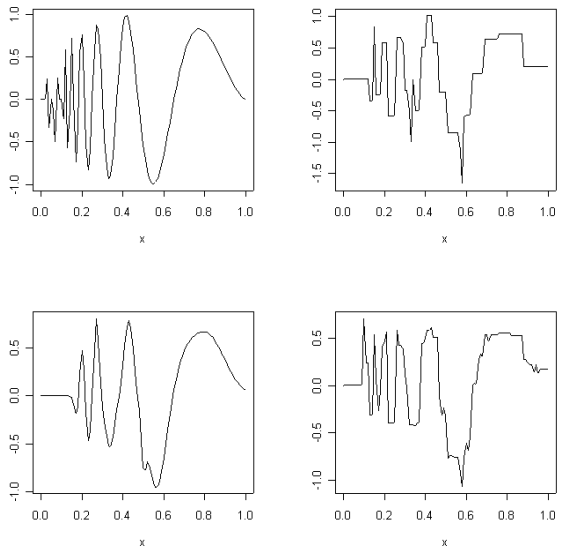

FIGURE 4. Experiment 1, $f = Doppler$ and $\sigma = 0.3$. Up-left: true regression function. Down-left: SVM. Up-right: wavelet estimate with our algorithm. Down-right: "classical" wavelet estimate.

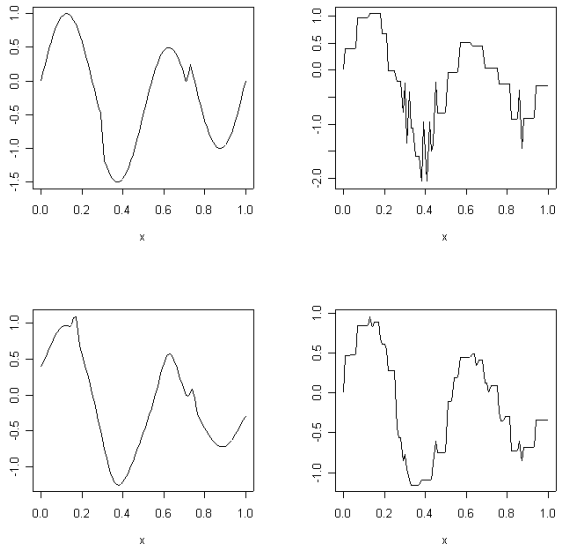

FIGURE 5. Experiment 2, $f = HeaviSine$ and $\sigma = 0.3$.

Definition 4.1. We put, for any $\theta \in \Theta$:

$$r_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - f_\theta(X_i)\right)^2$$

$$r_2(\theta) = \frac{1}{N} \sum_{i=N+1}^{2N} \left(Y_i - f_\theta(X_i)\right)^2 .$$

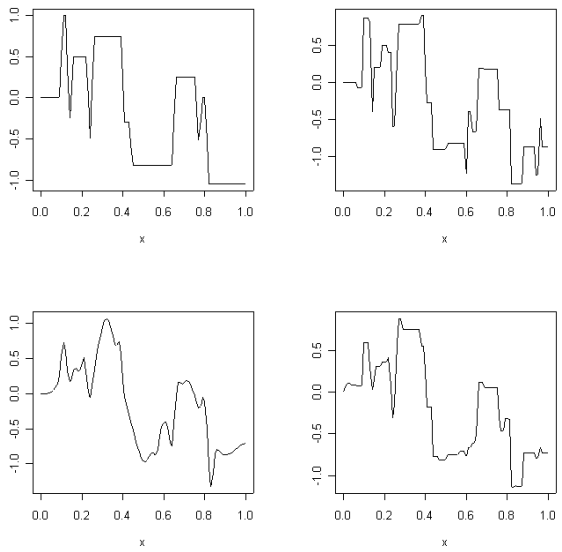

FIGURE 6. Experiment 3, $f = Blocks$ and $\sigma = 0.3$.

Our objective is:

$$\overline{\theta}_2 = \arg\min_{\theta\in\Theta} r_2(\theta),$$

if the minimum of r_2 is not unique then we take for $\overline{\theta}_2$ any element of Θ reaching the minimum value of r_2.

Let Θ_0 be a finite family of vectors belonging to Θ, so that $|\Theta_0| = m$. Actually, Θ_0 is allowed to be data-dependant:

$$\Theta_0 = \Theta_0(X_1, ..., X_{2N})$$

but we assume that the function $(X_1, ..., X_{2N}) \mapsto \Theta_0(X_1, ..., X_{2N})$ is exchangeable with respect to its $2N$ arguments, and is such that $m = m(N)$ depends only on N, not on $(X_1, ..., X_{2N})$.

The problem of the indexation of the elements of Θ_0 is not straightforward and we must be very careful about it. Let $<_\Theta$ be a complete order on Θ, and write:

$$\Theta_0 = \{\theta_1, ..., \theta_m\}$$

where

$$\theta_1 <_\Theta ... <_\Theta \theta_m.$$

Remark that, in this case, every θ_k is an exchangeable function of $(X_1, ..., X_{2N})$. In some cases, we will use other indexations. For example, in the case of SVM, we will take $m = 2N$ and:

$$\Theta_0 = \{\Psi(X_1), ..., \Psi(X_{2N})\}.$$

Clearly, there is no reason for having $\theta_1 = \Psi(X_1)$. In such a case, if necessary we can use another notation, for example define $\theta_i^* = \Psi(X_i)$. Then we will have:

$$\Theta_0 = \{\theta_1^*, ..., \theta_m^*\}$$

where θ_i^* is not an exchangeable function of $(X_1, ..., X_{2N})$.

Now, let us write, for any $k \in \{1, ..., m\}$:

$$\alpha_1^k = \arg\min_{\alpha \in \mathbb{R}} r_1(\alpha\theta_k) = \frac{\sum_{i=1}^N f_{\theta_k}(X_i)Y_i}{\sum_{i=1}^N f_{\theta_k}(X_i)^2}$$
$$\alpha_2^k = \arg\min_{\alpha \in \mathbb{R}} r_2(\alpha\theta_k) = \frac{\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)Y_i}{\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}$$
$$\mathcal{C}^k = \frac{\frac{1}{N}\sum_{i=1}^N f_{\theta_k}(X_i)^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}.$$

4.2. **Basics Results.**

Theorem 4.1. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1, ..., m\}$:*

$$r_2\left[(\mathcal{C}^k\alpha_1^k).\theta_k\right] - r_2(\alpha_2^k.\theta_k) \leq 4\left[\frac{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2Y_i^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}\right]\frac{\log\frac{2m}{\varepsilon}}{N}.$$

Remark 4.1. Here again, it is possible to make some hypothesis in order to make the right-hand side of the theorem observable. In particular, if we assume that:

$$\exists B \in \mathbb{R}_+, \quad P(|Y| \leq B) = 1,$$

then we can get a looser observable upper bound:

$$P_{2N}\Bigg\{\forall k \in \{1, ..., m\}, \quad r_2\left[(\mathcal{C}^k\alpha_1^k).\theta_k\right] - r_2(\alpha_2^k.\theta_k)$$
$$\leq 4\left[B^2 + \frac{\frac{1}{N}\sum_{i=1}^N f_{\theta_k}(X_i)^2Y_i^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}\right]\frac{\log\frac{2m}{\varepsilon}}{N}\Bigg\} \geq 1-\varepsilon.$$

If we don't want to make this assumption, we can use the following variant, that gives a first-order approximation for the bound.

Theorem 4.2. *For any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1, ..., m\}$:*

$$r_2\left[(\mathcal{C}^k\alpha_1^k).\theta_k\right] - r_2(\alpha_2^k.\theta_k)$$
$$\leq \frac{8\log\frac{4m}{\varepsilon}}{N}\left[\frac{\frac{1}{N}\sum_{i=1}^N f_{\theta_k}(X_i)^2Y_i^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2} + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4Y_i^4\log\frac{2m}{\varepsilon}}{2N}}\right].$$

Remark 4.2. Let us assume that Y is such that we know two constants b_Y and B_Y such that:

$$P\exp\left(b_Y Y\right) \leq B_Y < \infty.$$

Then we have, with probability at least $1-\varepsilon$:

$$\sup_{i\in\{1,...,2N\}} Y_i \leq \frac{1}{b_Y}\log\frac{2NB_Y}{\varepsilon}.$$

So the bound of the theorem leads to a looser observable bound:

$$r_2\left[(\mathcal{C}^k\alpha_1^k).\theta_k\right] - r_2(\alpha_2^k.\theta_k)$$
$$\leq \frac{8\log\frac{8m}{\varepsilon}}{N}\left[\frac{\frac{1}{N}\sum_{i=1}^N f_{\theta_k}(X_i)^2Y_i^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2} + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\log\frac{4m}{\varepsilon}\log^4\frac{4NB_Y}{\varepsilon}}{2Nb_Y^4}}\right].$$

A proof of this assertion is given in the next section.

The proofs of both theorems are given in the next section: however, we are going to see at first how to apply this result.

Let us compare the first order term of this theorem to the analogous term in the inductive case (theorems 2.1 and 2.2). The factor of the variance term is 8 instead of 2 in the inductive case. A factor 2 is to be lost because we have here the variance of a sample of size $2N$ instead of N in the inductive case. But another factor 2 is lost here. Moreover, in the inductive case, we had the real variance of $Yf(X)$ instead of the moment of order 2 here.

In the next subsection, we give several improvements of these bounds, that allows to recover a real variance, and to recover the factor 2. We also give a version that allows to deal with a test sample of different size, this being a generalization of theorem 4.1 more than of its improved variants.

4.3. **Improvements and generalization of the bound.** The proof of all the theorems of this subsection is given in the next section.

4.3.1. *Relative bounds.* We introduce some new notations.

Definition 4.2. We write:

$$\forall \theta \in \Theta, r_{1,2}(\theta) = r_1(\theta) + r_2(\theta)$$

and, in the case of a model $k \in \{1, ..., m\}$:

$$\alpha_{1,2}^k = \arg\min_{\alpha \in \mathbb{R}} r_{1,2}(\alpha\theta_k).$$

The we have the following theorem.

Theorem 4.3. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1, ..., m\}$:*

$$r_2(\mathcal{C}^k \alpha_1^k \theta_k) - r_2(\alpha_2^k \theta_k)) \leq 4 \left[\frac{\frac{1}{N}\sum_{i=1}^{2N} \left[f_{\theta_k}(X_i) Y_i - \alpha_{1,2}^k f_{\theta_k}(X_i)^2 \right]^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2} \right] \frac{\log \frac{2m}{\varepsilon}}{N}.$$

It is moreover possible to modify the upper bound to make it observable. We obtain that with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1, ..., m\}$:

$$\begin{aligned} &r_2\left[(\mathcal{C}^k \alpha_1^k)\theta_k\right] - r_2(\alpha_2^k \theta_k) \\ &\qquad \leq \frac{16 \log \frac{4m}{\varepsilon}}{N} \left[\frac{1}{N} \sum_{i=1}^{N} \left(f_{\theta_k}(X_i) Y_i - \alpha_1^k f_{\theta_k}(X_i)^2 \right)^2 \right] + \mathcal{O}\left(\left[\frac{\log \frac{m}{\varepsilon}}{N} \right]^{\frac{3}{2}} \right). \end{aligned}$$

So we can see that this theorem is an improvement on theorem 4.1 when some features $f_{\theta_k}(.)$ are well correlated with Y. But we loose another factor 2 by making the first-order term of the bound observable.

4.3.2. *Improvement of the variance term.*

Theorem 4.4. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1, ..., m\}$:*

$$r_2(\mathcal{C}^k \alpha_1^k \theta_k) - r_2(\alpha_2^k \theta_k)) \leq \left[\frac{1}{1 - \frac{2 \log \frac{2m}{\varepsilon}}{N}} \right] \frac{2 \log \frac{2m}{\varepsilon}}{N} \frac{V_1(\theta_k) + V_2(\theta_k)}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2},$$

where:

$$V_1(\theta_k) = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i f_{\theta_k}(X_i) - \frac{1}{N}\sum_{j=1}^{N} Y_j f_{\theta_k}(X_j)\right]^2,$$

$$V_2(\theta_k) = \frac{1}{N}\sum_{i=N+1}^{2N}\left[Y_i f_{\theta_k}(X_i) - \frac{1}{N}\sum_{j=N+1}^{2N} Y_j f_{\theta_k}(X_j)\right]^2.$$

It is moreover possible to give an observable upper bound: we obtain that with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1,...,m\}$:

$$r_2\left[(\mathcal{C}^k\alpha_1^k)\theta_k\right] - r_2(\alpha_2^k\theta_k) \leq \left[\frac{1}{1-\frac{2\log\frac{4m}{\varepsilon}}{N}}\right]\frac{4\log\frac{4m}{\varepsilon}}{N}\frac{V_1(\theta_k)}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}$$
$$+\left[\frac{1}{1-\frac{2\log\frac{4m}{\varepsilon}}{N}}\right]2\left(2+\sqrt{2}\right)\left(\frac{\log\frac{6m}{\varepsilon}}{N}\right)^{\frac{3}{2}}\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4 Y_i^4}}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}.$$

Here again, we can make the bound fully observable under an exponential moment assumption about Y.

4.3.3. *Test sample of different size.* In the context of classification, Catoni [6] gave a method in order to be able to deal with the case where the test sample is of size $(k+1)N$ where k is an integer greater than 0. More precisely, we assume that $P_{(k+1)N}$ is an exchangeable probability distribution on $(\mathcal{X}\times\mathbb{R})^{(k+1)N}$ and that we observe:

$$(X_1,Y_1),...,(X_N,Y_N) \quad \text{and} \quad X_{N+1},...,X_{(k+1)N}.$$

In the case where $k>1$, the variance term will be better than in the case where $k=1$. This method can be used in the setting of regression too.

Definition 4.3. From now, we will use the notation, when $k \neq 1$:

$$r_1(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - f_\theta(X_i)\right)^2$$

$$r_2(\theta) = \frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\left(Y_i - f_\theta(X_i)\right)^2.$$

We still consider a family:

$$\Theta_0(X_1,...,X_{(k+1)N}) = \{\theta_1,...,\theta_m\}$$

that is data-dependent in an exchangeable way, with the same indexation convention than in the case where $k=1$. Now, let us write, for any $h \in \{1,...,m\}$:

$$\alpha_1^h = \arg\min_{\alpha\in\mathbb{R}} r_1(\alpha\theta_h) = \frac{\sum_{i=1}^{N} f_{\theta_h}(X_i)Y_i}{\sum_{i=1}^{N} f_{\theta_h}(X_i)^2}$$

$$\alpha_2^h = \arg\min_{\alpha\in\mathbb{R}} r_2(\alpha\theta_h) = \frac{\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)Y_i}{\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)^2}$$

$$\mathcal{C}^h = \frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_h}(X_i)^2}{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)^2}.$$

Let us finally put:

$$\mathbf{P} = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\delta_{Z_i},$$

and, for any $\theta \in \Theta$:

$$\mathbb{V}_\theta = \mathbf{P}\left\{ \left[\Big(f_\theta(X)Y \Big) - \mathbf{P}\Big(f_\theta(X)Y \Big) \right]^2 \right\}.$$

Then we have the following theorem.

Theorem 4.5. *Let us assume that we have constants B_h and β_h such that, for any $h \in \{1, ..., m\}$:*

$$P \exp\left(\beta_h \left| f_{\theta_h}(X_i) Y_i \right| \right) \leq B_h.$$

For any $\varepsilon > 0$, with $P_{(k+1)N}$ probability at least $1-\varepsilon$ we have, for any $h \in \{1, ..., m\}$:

$$r_2\left(\mathcal{C}^h \alpha_1^h \theta_h\right) - r_2\left(\alpha_2^h \theta_h\right) \leq \frac{\left(1+\frac{1}{k}\right)^2}{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)^2} \left[\frac{2\mathbb{V}_{\theta_h} \log \frac{4m}{\varepsilon}}{N} + \frac{16\left(\log\frac{4m}{\varepsilon}\right)^{\frac{3}{2}} \left(\log \frac{4(k+1)mNB_h}{\varepsilon}\right)^3}{3\beta_h^3 N^{\frac{3}{2}} \mathbb{V}_{\theta_h}^{\frac{1}{2}}} + \frac{64\left(\log\frac{4m}{\varepsilon}\right)^2 \left(\log \frac{4(k+1)mNB_h}{\varepsilon}\right)^6}{9\beta_h^6 N^2 \mathbb{V}_{\theta_h}^2} \right].$$

Here again, it is possible to replace the variance term by its natural estimator:

$$\hat{\mathbb{V}}_{\theta_h} = \frac{1}{N}\sum_{i=1}^{N} \left[f_\theta(X_i)Y_i - \frac{1}{N}\sum_{j=1}^{N} f_\theta(X_j)Y_j \right]^2.$$

4.4. **Application to regression estimation.** We give here the interpretation of the preceding theorems in terms of confidence; this motivates an algorithm similar to the one described in the inductive case.

Definition 4.4. We take, for any $(\theta, \theta') \in \Theta^2$:

$$d_2(\theta, \theta') = \sqrt{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} \left[f_\theta(X_i) - f_{\theta'}(X_i) \right]^2} = \sqrt{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} \langle \theta - \theta', \Psi(X_i) \rangle^2}.$$

Let also $\|\theta\|_2 = d_2(\theta, 0)$ and:

$$\langle \theta, \theta' \rangle_2 = \frac{1}{(k+1)N} \sum_{i=N+1}^{(k+1)N} f_\theta(X_i) f_{\theta'}(X_i).$$

We define, for any $h \in \{1, ..., m\}$ and ε:

$$\mathcal{CR}(h, \varepsilon) = \left\{ \theta \in \Theta : \left| \left\langle \theta - \mathcal{C}^h \alpha_1^h \theta_h, \theta_h \right\rangle_2 \right| \leq \sqrt{\beta(\varepsilon, h)} \right\},$$

where $\beta(\varepsilon, h)$ is the upper bound in theorem 4.1 (or in the other theorems given previously).

For the same reasons as in the inductive case, these theorems implies the following result.

Corollary 4.6. *We have:*

$$P_{2N}\left[\forall h \in \{1, ..., m\}, \overline{\theta}_2 \in \mathcal{CR}(h, \varepsilon) \right] \geq 1 - \varepsilon.$$

Definition 4.5. We call $\Pi_2^{h,\varepsilon}$ the orthogonal projection into $\mathcal{CR}(h, \varepsilon)$ with respect to the distance d_2.

We propose the following algorithm:

- choose $\theta(0) \in \Theta$ (for example 0);

FIGURE 7. Detailed version of the feature selection algorithm in the transductive case.

We have $\varepsilon > 0$, $\kappa > 0$, N observations $(X_1, Y_1), ..., (X_N, Y_N)$ and also $X_{N+1}, ..., X_{(k+1)N}$, m features $f_{\theta_1}(.), ..., f_{\theta_m}(.)$ and $\theta(0) = (\theta_1(0), ..., \theta_m(0)) = (0, ..., 0) \in \mathbb{R}^m$. In a first time, compute every α_1^h and $\beta(\varepsilon, h)$ for $h \in \{1, ..., m\}$. Set $n \leftarrow 0$.

Repeat:

- set $n \leftarrow n + 1$;
- set *best_improvement* $\leftarrow 0$ and $\theta(n) = \theta(n-1)$;
- for $h \in \{1, ..., m\}$, compute:

$$v_h = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)^2,$$

$$\gamma_h \leftarrow \alpha_1^h - \frac{1}{v_h} \sum_{j=1}^{m} \theta_j(n) \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} f_{\theta_j}(X_i) f_{\theta_h}(X_i),$$

$$\delta_h \leftarrow v_h \Big(|\gamma_h| - \beta(\varepsilon, h) \Big)_+^2,$$

  and if $\delta_h >$ *best_improvement*, set:

$$\textit{best\_improvement} \leftarrow \delta_h,$$

$$h(n) \leftarrow h;$$

- if *best_improvement* > 0:

$$\theta_{h(n)}(n) \leftarrow \theta_{h(n)}(n) + sgn(\gamma_h) \Big(|\gamma_h| - \beta(\varepsilon, h) \Big)_+;$$

until *best_improvement* $< \kappa$.

Return the estimation:

$$\Big[\tilde{Y}_{N+1}, ..., \tilde{Y}_{(k+1)N} \Big] = \Big[\hat{f}(X_{N+1}), ..., \hat{f}(X_{(k+1)N}) \Big]$$

where:

$$\hat{f}(.) = \sum_{h=1}^{m} \theta_h(n) f_{\theta_h}(.).$$

- at step $n \in \mathbb{N}^*$, we have: $\theta(0), ..., \theta(n-1)$. Choose $h(n)$, for example:

$$h(n) = \arg \max_{h \in \{1, ..., m\}} d_2(\theta(n-1), \mathcal{CR}(h, \varepsilon)),$$

  and take:

$$\theta(n) = \Pi_2^{h(n), \varepsilon} \theta(n-1);$$

- we can use the following stopping rule: $\|\theta(n-1) - \theta(n)\|_2^2 \leq \kappa$ where $0 < \kappa < \frac{1}{N}$.

Definition 4.6. We write n_0 the stopping step, and:

$$\hat{f}(.) = f_{\theta(n_0)}(.)$$

the corresponding function.

Here again we give a detailed version of the algorithm, see figure 7. Remark that as in the inductive case, we are allowed to use whatever heuristic to choose $k(n)$ if we want to avoid the maximization.

Theorem 4.7. *We have:*

$$P_{2N}\left[\forall n \in \{1,...,n_0\}, r_2\left[\theta\left(n\right)\right] \leq r_2\left[\theta\left(n-1\right)\right] - d_2^2\left[\theta(n),\theta(n-1)\right]\right] \geq 1-\varepsilon$$

The proof of this theorem is exactly the same as the proof of theorem 2.4.

Example 4.1 (Estimation of wavelet coefficients)**.** Let us consider the case where Θ_0 does not depend on the observations. We can, for example, choose a basis of Θ, or a basis of a subspace of Θ. We obtain an estimator of the form:

$$\hat{f}(x) = \sum_{h=1}^{m} \alpha^h f_{\theta_h}(x).$$

In the case when $(f_{\theta_k})_k$ is a wavelet basis, then we obtain here again a procedure for thresholding wavelets coefficients.

Example 4.2 (SVM and multiscale SVM)**.** Let us choose Θ as the set of all functions $\mathcal{X} \to \mathbb{R}$, $f_\theta(x) = \theta(x)$, a family of kernels $K_1, ..., K_{m'(N)}$ for a $m'(N) \geq 1$ and:

$$\Theta_0 = \left\{K_h(X_i,.), h \in \{1,...,m'(N)\}, i \in \{1,...,(k+1)N\}\right\}.$$

In this case we have $m = (k+1)Nm'(N)$. We obtain an estimator of the form:

$$\hat{f}(x) = \sum_{h=1}^{m'(N)} \sum_{j=1}^{2N} \alpha^{j,h} K_h(X_j, x).$$

Let us put:

$$I_h = \left\{j \in \{1,...,2N\}, \alpha^{j,h} \neq 0\right\}.$$

We have:

$$\hat{f}(x) = \sum_{h=1}^{m'(N)} \sum_{j \in I_h} \alpha^{j,h} K_h(X_i, x),$$

that is a Support Vector Machine with different kernels estimate; like in example 2.3, the kernels K_h can be the same kernel taken at different scales.

Example 4.3 (Kernel PCA Kernel Projection Machine)**.** Let us take Θ as a Hilbert space, with scalar product $\langle .,. \rangle$, let us take a function $\Psi : \mathcal{X} \to \Theta$ and consider the kernel:

$$K(x,x') = \left\langle \Psi(x), \Psi(x') \right\rangle.$$

Let us consider a principal component analysis (PCA) of the family:

$$\left\{\Psi(X_1), ..., \Psi(X_{(k+1)N})\right\}$$

by performing a diagonalization of the matrix:

$$\left(K\left(X_i, X_j\right)\right)_{1 \leq i,j \leq (k+1)N}.$$

This method is known as Kernel PCA, see for example Schlkopf, Smola and Mller [13]. We obtain eigenvalues:

$$\lambda^1 \geq ... \geq \lambda^{(k+1)N}$$

and associated eigenvectors $e^1, ..., e^{(k+1)N}$, associated to elements of Θ:

$$\Psi_1 = \sum_{i=1}^{(k+1)N} e_i^1 \Psi(X_i), ..., \Psi_{(k+1)N} = \sum_{i=1}^{(k+1)N} e_i^{(k+1)N} \Psi(X_i)$$

that are exchangeable functions of the observations. Using the family:

$$\Theta_0 = \left\{\Psi_1, ..., \Psi_{(k+1)N}\right\}$$

we obtain an algorithm that selects which eigenvectors are going to be used in the regression estimation. This is very close to the Kernel Projection Machine (KPM) described by Blanchard, Massart, Vert and Zwald [1] in the context of classification.

5. Proof of the theorems in the transductive case

5.1. **Proof of theorems 4.1 and 4.2.** Here again, the first thing to do is to prove a general deviation inequality. This one is a variant of the one given by Catoni [5]. We go back to the notations oh theorem 4.1 and 4.2, with test sample of size N.

Definition 5.1. Let $\mathcal{G}$ denote the set of all functions:

$$\begin{aligned} g : (\mathcal{X} \times \mathbb{R})^{2N} \times \mathbb{R}^2 &\to \mathbb{R} \\ (Z_1, ..., Z_{2N}, u, u') &\mapsto g(Z_1, ..., Z_{2N}, u, u') = g(u, u') \end{aligned}$$

for the sake of simplicity, such that g is exchangeable with respect to its $2N$ first arguments.

Lemma 5.1. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1, ..., Z_{2N})$, for any measurable function $\eta : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any measurable function $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\theta \in \Theta$ and any $g \in \mathcal{G}$:*

$$\mathcal{P} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \left\{ g\left[f_\theta(X_{i+N}), Y_{i+N} \right] - g\left[f_\theta(X_i), Y_i \right] \right\} - \frac{\lambda^2}{c_g N^2} \sum_{i=1}^{2N} g\left[f_\theta(X_i), Y_i \right]^2 - \eta \right) \leq \mathcal{P} \exp\left(-\eta\right)$$

and the reverse inequality:

$$\mathcal{P} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \left\{ g\left[f_\theta(X_i), Y_i \right] - g\left[f_\theta(X_{i+N}), Y_{i+N} \right] \right\} - \frac{\lambda^2}{c_g N^2} \sum_{i=1}^{2N} g\left[f_\theta(X_i), Y_i \right]^2 - \eta \right) \leq \mathcal{P} \exp\left(-\eta\right),$$

where we write:

$$\begin{aligned} \eta &= \eta\left((X_1, Y_1), ..., (X_{2N}, Y_{2N})\right) \\ \lambda &= \lambda\left((X_1, Y_1), ..., (X_{2N}, Y_{2N})\right) \end{aligned}$$

for short, and:

$$c_g = \begin{cases} 2 & \text{if } g \text{ is nonnegative,} \\ 1 & \text{otherwise.} \end{cases}$$

Proof. In order to prove the first inequality, we write:

$$\mathcal{P}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\Big\{g\Big[f_\theta(X_{i+N}),Y_{i+N}\Big]-g\Big[f_\theta(X_i),Y_i\Big]\Big\} - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N} g\Big[f_\theta(X_i),Y_i\Big]^2-\eta\right)$$
$$=\mathcal{P}\exp\left(\sum_{i=1}^{N}\log\cosh\left\{\frac{\lambda}{N}g\Big[f_\theta(X_{i+N}),Y_{i+N}\Big]-\frac{\lambda}{N}g\Big[f_\theta(X_i),Y_i\Big]\right\} - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N} g\Big[f_\theta(X_i),Y_i\Big]^2-\eta\right).$$

This last step is true because $\mathcal{P}$ is exchangeable. We conclude by using the inequality:

$$\forall x\in\mathbb{R}, \log\cosh x\leq\frac{x^2}{2}.$$

We obtain:

$$\log\cosh\left\{\frac{\lambda}{N}g\Big[f_\theta(X_{i+N}),Y_{i+N}\Big]-\frac{\lambda}{N}g\Big[f_\theta(X_i),Y_i\Big]\right\}$$
$$\leq\frac{\lambda^2}{2N^2}\left\{g\Big[f_\theta(X_{i+N}),Y_{i+N}\Big]-g\Big[f_\theta(X_i),Y_i\Big]\right\}^2\leq\frac{\lambda^2}{c_gN^2}g\Big[f_\theta(X_i),Y_i\Big]^2.$$

The proof for the reverse inequality is exactly the same. □

We can now give the proof of the theorems.

Proof of theorem 4.1. From now we assume that the hypothesis of theorem 4.1 are satisfied. Let us choose $\varepsilon'>0$ and apply lemma 5.1 with $\eta=-\log\varepsilon'$, and g such that $g(u,u')=uu'$. We obtain: for any exchangeable distribution $\mathcal{P}$, for any measurable function $\lambda:(\mathcal{X}\times\mathbb{R})^{2N}\to\mathbb{R}_+^*$ that is exchangeable with respect to its $2\times 2N$ arguments, for any $\theta\in\Theta$:

$$\mathcal{P}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\Big[f_\theta(X_{i+N})Y_{i+N}-f_\theta(X_i)Y_i\Big]-\frac{\lambda^2}{N^2}\sum_{i=1}^{2N}f_\theta(X_i)^2Y_i^2+\log\varepsilon'\right)\leq\varepsilon'$$

and the reverse inequality:

$$\mathcal{P}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\Big[f_\theta(X_i)Y_i-f_\theta(X_{i+N})Y_{i+N}\Big]-\frac{\lambda^2}{N^2}\sum_{i=1}^{2N}f_\theta(X_i)^2Y_i^2+\log\varepsilon'\right)\leq\varepsilon'.$$

Let us denote:

$$f(\theta,\varepsilon',\lambda)=\lambda\left|\frac{1}{N}\sum_{i=1}^{N}\Big[f_\theta(X_{i+N})Y_{i+N}-f_\theta(X_i)Y_i\Big]-\frac{\lambda^2}{N^2}\sum_{i=1}^{2N}f_\theta(X_i)^2Y_i^2\right|+\log\varepsilon'.$$

The previous inequalities imply that: for any exchangeable $\mathcal{P}$, for any measurable function $\lambda:(\mathcal{X}\times\mathbb{R})^{2N}\to\mathbb{R}_+^*$ that is exchangeable with respect to its $2\times 2N$ arguments, for any $\theta\in\Theta$:

$$\mathcal{P}\exp f((Z_1,...,Z_{2N}),\theta,\varepsilon',\lambda)\leq 2\varepsilon'. \tag{5.1}$$

Now, let us introduce a new conditional probability measure:

$$\overline{P}=\frac{1}{(2N)!}\sum_{\sigma\in\mathfrak{S}_{2N}}\delta_{(X_{\sigma_i},Y_{\sigma_i})_{i\in\{1,\dots,2N\}}}.$$

Remark that P_{2N} being exchangeable, we have, for any bounded function $h : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$,

$$P_{2N} h = P_{2N}\left(\overline{P} h\right).$$

The measure $\overline{P}$ is exchangeable, so we can apply equation 5.1. For any values of $Z_1, ..., Z_{2N}$ we have:

$$\forall \theta \in \Theta, \quad \overline{P} \exp f((Z_1, ..., Z_{2N}), \theta, \varepsilon', \lambda) \leq 2\varepsilon'.$$

In particular, we can choose $\theta = \theta(Z_1, ..., Z_{2N})$ as an exchangeable function of $(Z_1, ..., Z_{2N})$, because we will have:

$$\frac{1}{(2N)!} \sum_{\sigma \in \mathfrak{S}_{2N}} \exp f(Z_{\sigma(1)}, ..., Z_{\sigma(2N)}), \theta(Z_{\sigma(1)}, ..., Z_{\sigma(2N)})), \varepsilon', \lambda)$$
$$= \frac{1}{(2N)!} \sum_{\sigma \in \mathfrak{S}_{2N}} \exp f(Z_{\sigma(1)}, ..., Z_{\sigma(2N)}), \theta(Z_1, ..., Z_{2N})), \varepsilon', \lambda) \leq \varepsilon'.$$

Here, we choose as functions θ the members of Θ_0: $\theta_1, ..., \theta_m$ (remember that we choose this indexation in such a way that for any k, θ_k is an exchangeable function of $(Z_1, ..., Z_{2N})$). We have, for any $\lambda_1, ..., \lambda_m$ that are m exchangeable functions of $(Z_1, ..., Z_{2N})$:

$$P_{2N}\left[\exists k \in \{1, ..., m\}, f((Z_1, ..., Z_{2N}), \theta_k, \varepsilon', \lambda_k) > 0\right]$$
$$= P_{2N}\left[\bigcup_{k=1}^{m} \{f((Z_1, ..., Z_{2N}), \theta_k, \varepsilon', \lambda_k) > 0\}\right]$$
$$\leq P_{2N}\left[\sum_{k=1}^{m} \mathbb{1}\left(f((Z_1, ..., Z_{2N}), \theta_k, \varepsilon', \lambda_k) > 0\right)\right]$$
$$= P_{2N}\overline{P}\left[\sum_{k=1}^{m} \mathbb{1}\left(f((Z_1, ..., Z_{2N}), \theta_k, \varepsilon', \lambda_k) > 0\right)\right]$$
$$= P_{2N} \sum_{k=1}^{m} \overline{P}\Big[\mathbb{1}\left(f((Z_1, ..., Z_{2N}), \theta_k, \varepsilon', \lambda_k) > 0\right)\Big]$$
$$\leq P_{2N} \sum_{k=1}^{m} \overline{P} \exp f((Z_1, ..., Z_{2N}), \theta_k, \varepsilon', \lambda_k).$$

Now let us apply inequality 5.1, we obtain:

$$P_{2N}\left[\exists k \in \{1, ..., m\}, f((Z_1, ..., Z_{2N}), \theta_k, \varepsilon', \lambda_k) > 0\right] \leq P_{2N} \sum_{k=1}^{m} 2\varepsilon' = 2\varepsilon' m = \varepsilon$$

if we choose:

$$\varepsilon' = \frac{\varepsilon}{2m}.$$

From now, we assume that the event:

$$\left\{\forall k \in \{1, ..., m\}, f\left((Z_1, ..., Z_{2N}), \theta_k, \frac{\varepsilon}{2m}, \lambda_k\right) \leq 0\right\}$$

is satisfied. It can be written, for any $k \in \{1, ..., m\}$:

$$\left|\frac{1}{N} \sum_{i=1}^{N}\Big[f_{\theta_k}(X_{i+N}) Y_{i+N} - f_{\theta_k}(X_i) Y_i\Big]\right| \leq \frac{\lambda_k}{N^2} \sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2 + \frac{\log \frac{2m}{\varepsilon}}{\lambda_k}.$$

Let us divide both inequalities by:

$$\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2.$$

We obtain, for any $k \in \{1, ..., m\}$:

$$\left|\alpha_2^k - \mathcal{C}^k\alpha_1^k\right| \leq \frac{\frac{\lambda_k}{N^2}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2 + \frac{\log\frac{2m}{\varepsilon}}{\lambda_k}}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}.$$

It is now time to choose the functions λ_k. We try to optimize the right-hand side with respect to λ_k, and obtain a minimal value for:

$$\lambda_k = \sqrt{\frac{N\log\frac{2m}{\varepsilon}}{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2}}.$$

This choice is admissible because it is exchangeable with respect to $(Z_1, ..., Z_{2N})$.

So we have, for any $k \in \{1, ..., m\}$:

$$\left|\mathcal{C}^k\alpha_1^k - \alpha_2^k\right| \leq 2\frac{\sqrt{\frac{1}{N^2}\sum_{i=1}^{2N}\left[f_{\theta_k}(X_i)^2 Y_i^2\right]\log\frac{2m}{\varepsilon}}}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}.$$

Finally, remark that:

$$\left|\mathcal{C}^k\alpha_1^k - \alpha_2^k\right| = \sqrt{\frac{r_2\left[(\mathcal{C}^k\alpha_1^k)\theta_k\right] - r_2(\alpha_2^k\theta_k)}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}},$$

that leads to the conclusion that for any $k \in \{1, ..., m\}$:

$$r_2\left[(\mathcal{C}^k\alpha_1^k)\theta_k\right] - r_2(\alpha_2^k\theta_k) \leq 2^2\frac{\frac{1}{N^2}\sum_{i=1}^{2N}\left[f_{\theta_k}(X_i)^2 Y_i^2\right]\log\frac{2m}{\varepsilon}}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}.$$

This ends the proof. □

Proof of theorem 4.2. We write:

$$\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2 = \frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2 Y_i^2 + \frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2$$

and try to upper bound the second term. We apply lemma 5.1, but this time with g such that $g(u) = (uu')^2$ that is nonnegative, and obtain, for any ε, for any (exchangeables) θ and λ:

$$\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2 \leq \frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2 Y_i^2 + \frac{\lambda}{2N}\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4 Y_i^4 + \frac{\log\varepsilon}{\lambda}.$$

We choose:

$$\lambda = \sqrt{\frac{2N\log\varepsilon}{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4 Y_i^4}},$$

we apply this result to every $\theta \in \Theta_0$, and combine it with theorem 4.1 by a union bound argument to obtain the result. □

5.2. **Proof of theorem 4.3.** First of all, we give the following obvious variant of lemma 5.1:

Lemma 5.2. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1, ..., Z_{2N})$, for any measurable function $\eta : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any measurable function $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\theta \in \Theta$:*

$$\mathcal{P} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \left\{ \left[f_\theta(X_{i+N}) Y_{i+N} - \alpha(\theta) f_\theta(X_{i+N})^2 \right] - \left[f_\theta(X_i) Y_i - \alpha(\theta) f_\theta(X_i)^2 \right] \right\} \right.$$
$$\left. - \frac{\lambda^2}{N^2} \sum_{i=1}^{2N} \left[f_\theta(X_i) Y_i - \alpha(\theta) f_\theta(X_i)^2 \right]^2 - \eta \right) \leq \mathcal{P} \exp\left(-\eta\right)$$

and the reverse inequality, where:

$$\alpha(\theta) = \arg\min_{\alpha \in \mathbb{R}} r_{1,2}(\alpha\theta).$$

Proof. This is actually just an applicatin of lemma 5.1, we just need to remark that $\alpha(\theta)$ is an exchangeable function of $(Z_1, ..., Z_{2N})$, and so we can take in lemma 5.1:

$$g(u, u') = uu' - u^2 \alpha(\theta),$$

that means that:

$$g\left[f_\theta(X_i), Y_i \right] = f_\theta(X_i) Y_i - \alpha(\theta) f_\theta(X_i)^2.$$

□

Proof of theorem 4.3. Proceeding exactly in the same way as in the proof of theorem 4.1, we obtain the following inequality with probability at least $1 - \varepsilon$:

$$r_2(\mathcal{C}^k \alpha_1^k \theta_k) - r_2(\alpha_2^k \theta_k)) \leq 4 \left[\frac{\frac{1}{N} \sum_{i=1}^{2N} \left[f_{\theta_k}(X_i) Y_i - \alpha_{1,2}^k f_{\theta_k}(X_i)^2 \right]^2}{\frac{1}{N} \sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2} \right] \frac{\log \frac{2m}{\varepsilon}}{N}. \tag{5.2}$$

This proves the theorem. □

Before giving the proof of the next theorem, let us see how we can make the first order term observable in this theorem. For example, we can write:

$$\left[f_{\theta_k}(X_i) Y_i - \alpha_{1,2}^k f_{\theta_k}(X_i)^2 \right]^2$$
$$= \left[f_{\theta_k}(X_i) Y_i - \alpha_1^k f_{\theta_k}(X_i)^2 \right]^2 + \left[\alpha_1^k - \alpha_{1,2}^k \right]^2 f_{\theta_k}(X_i)^4$$
$$+ 2 \left[f_{\theta_k}(X_i) Y_i - \alpha_1^k f_{\theta_k}(X_i)^2 \right] \left[\alpha_1^k - \alpha_{1,2}^k \right] f_{\theta_k}(X_i)^2.$$

Remark that it is obvious that:

$$|\alpha_1^k - \alpha_{1,2}^k| \leq |\alpha_1^k - \alpha_2^k|,$$

and so:

$$\left[f_{\theta_k}(X_i) Y_i - \alpha_{1,2}^k f_{\theta_k}(X_i)^2 \right]^2$$
$$\leq \left[f_{\theta_k}(X_i) Y_i - \alpha_1^k f_{\theta_k}(X_i)^2 \right]^2 + \left[\alpha_1^k - \alpha_2^k \right]^2 f_{\theta_k}(X_i)^4$$
$$+ 2 \left| f_{\theta_k}(X_i) Y_i - \alpha_1^k f_{\theta_k}(X_i)^2 \right| \left| \alpha_1^k - \alpha_2^k \right| f_{\theta_k}(X_i)^2.$$

Now, just write:

$$\alpha_1^k - \alpha_2^k = \left(1 - \mathcal{C}^k\right) \alpha_1^k - \left(\mathcal{C}^k \alpha_1^k - \alpha_2^k\right)$$

and so we get:

$$\begin{aligned}\Big[f_{\theta_k}(X_i)Y_i - \alpha^k_{1,2} f_{\theta_k}(X_i)^2\Big]^2 \\ \leq \Big[f_{\theta_k}(X_i)Y_i - \alpha^k_1 f_{\theta_k}(X_i)^2\Big]^2 + \Big[\mathcal{C}^k\alpha^k_1 - \alpha^k_2\Big]^2 f_{\theta_k}(X_i)^4 \\ + 2\Big|\mathcal{C}^k\alpha^k_1 - \alpha^k_2\Big|\Big|(1-\mathcal{C}^k)\alpha^k_1\Big| f_{\theta_k}(X_i)^4 + \big(1-\mathcal{C}^k\big)^2\big(\alpha^k_1\big)^2 f_{\theta_k}(X_i)^4 \\ + 2\Big|f_{\theta_k}(X_i)Y_i - \alpha^k_1 f_{\theta_k}(X_i)^2\Big|\Big|\mathcal{C}^k\alpha^k_1 - \alpha^k_2\Big| f_{\theta_k}(X_i)^2 \\ + 2\Big|f_{\theta_k}(X_i)Y_i - \alpha^k_1 f_{\theta_k}(X_i)^2\Big|\Big|(\mathcal{C}^k-1)\alpha^k_1\Big| f_{\theta_k}(X_i)^2.\end{aligned}$$

So finally, equation 5.2 left us with a second degree inequality with respect to $\big|\mathcal{C}^k\alpha^k_1 - \alpha^k_2\big|$ or $r_2(\mathcal{C}^k\alpha^k_1\theta_k) - r_2(\alpha^k_2\theta_k)$) that we can solve to obtain the following result: with probability at least $1-\varepsilon$, as soon as we have:

$$\left[\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2\right]^2 > \left[\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\right]\frac{4\log\frac{2m}{\varepsilon}}{N},$$

which is always true for large enough N, the quantity $\big|\mathcal{C}^k\alpha^k_1 - \alpha^k_2\big|$ belongs to the interval:

$$\left[\frac{2\log\frac{2m}{\varepsilon}}{N}\frac{b \pm \sqrt{b^2 + a\left(\frac{N}{\log\frac{2m}{\varepsilon}}\left[\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2\right]^2 - \frac{4}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\right)}}{\left[\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2\right]^2 - \frac{4\log\frac{2m}{\varepsilon}}{N}\left[\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\right]}\right]$$

with the following notations:

$$\begin{aligned}a &= \frac{1}{N}\sum_{i=1}^{2N}\Big[\big|f_{\theta_k}(X_i)Y_i - \alpha^k_1 f_{\theta_k}(X_i)^2\big| + \big|\alpha^1_k(1-\mathcal{C}^k)\big| f_{\theta_k}(X_i)^2\Big]^2, \\ b &= \frac{1}{N}\sum_{i=1}^{2N} 2 f_{\theta_k}(X_i)^2\Big[\big|\alpha^1_k(1-\mathcal{C}^k)\big| f_{\theta_k}(X_i)^2 + \big|f_{\theta_k}(X_i)Y_i - \alpha^k_1 f_{\theta_k}(X_i)^2\big|\Big].\end{aligned}$$

Remark that only one of the bounds of the interval is positive. So we obtain the following result: with P_{2N}-probability at least $1-\varepsilon$, as soon as:

$$\left[\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2\right]^2 > \left[\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\right]\frac{4\log\frac{2m}{\varepsilon}}{N}$$

we have:

$$\forall k \in \{1,...,m\}, \quad r_2\left[(\mathcal{C}^k\alpha^k_1)\theta_k\right] - r_2(\alpha^k_2\theta_k) \leq \frac{4\log^2\frac{2m}{\varepsilon}}{N^2}\left[\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2\right]$$
$$\left[\frac{b + \sqrt{b^2 + a\left(\frac{N}{\log\frac{2m}{\varepsilon}}\left[\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2\right]^2 - \frac{4}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\right)}}{\left[\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2\right]^2 - \frac{4\log\frac{2m}{\varepsilon}}{N}\left[\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\right]}\right]^2.$$

We can notice that this bound may be written:

$$r_2\left[(\mathcal{C}^k\alpha_1^k)\theta_k\right] - r_2(\alpha_2^k\theta_k) \leq \frac{8a\log\frac{2m}{\varepsilon}}{N} + \mathcal{O}\left(\left[\frac{\log\frac{m}{\varepsilon}}{N}\right]^{\frac{3}{2}}\right)$$
$$= \frac{8\log\frac{2m}{\varepsilon}}{N}\left[\frac{1}{N}\sum_{i=1}^{2N}\left(f_{\theta_k}(X_i)Y_i - \alpha_1^k f_{\theta_k}(X_i)^2\right)^2\right] + \mathcal{O}\left(\left[\frac{\log\frac{m}{\varepsilon}}{N}\right]^{\frac{3}{2}}\right).$$

The next step would be now to replace the bound by an observable quantity, by getting a bound like:

$$\frac{1}{N}\sum_{i=1}^{2N}\left(f_{\theta_k}(X_i)Y_i - \alpha_1^k f_{\theta_k}(X_i)^2\right)^2$$
$$\leq \frac{2}{N}\sum_{i=1}^{N}\left(f_{\theta_k}(X_i)Y_i - \alpha_1^k f_{\theta_k}(X_i)^2\right)^2 + \mathcal{O}\left(\frac{\log\frac{m}{\varepsilon}}{N}\right)$$

with high probability. This can be done very simply, using lemma 5.1 with this time:

$$g(u,u') = \left(uu' - u^2\alpha(\theta)\right)^2.$$

We obtain the bound:

$$r_2\left[(\mathcal{C}^k\alpha_1^k)\theta_k\right] - r_2(\alpha_2^k\theta_k)$$
$$\leq \frac{16\log\frac{4m}{\varepsilon}}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\left(f_{\theta_k}(X_i)Y_i - \alpha_1^k f_{\theta_k}(X_i)^2\right)^2\right] + \mathcal{O}\left(\left[\frac{\log\frac{m}{\varepsilon}}{N}\right]^{\frac{3}{2}}\right).$$

5.3. **Proof of theorem 4.4.** The proof is exactly similar, we just use a new variant of lemma 5.1, that is based on an idea introduced by Catoni [6] in the context of classification.

Definition 5.2. Let us write:

$$T_\theta(Z_i) = f_\theta(X_i)Y_i$$

for short. We also introduce a conditional probability measure:

$$\mathcal{P}^{(2)} = \frac{1}{N!}\sum_{\sigma\in\mathfrak{S}_N}\delta_{(Z_1,\ldots,Z_N,Z_{N+\sigma(1)},\ldots,Z_{N+\sigma(N)})}.$$

Remark that, because $\mathcal{P}$ is exchangeable, we have, for any function h:

$$\mathcal{P}h = \mathcal{P}\left[\mathcal{P}^{(2)}h\right].$$

Lemma 5.3. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1,...,Z_{2N})$, for any measurable function $\eta : (\mathcal{X}\times\mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2\times 2N$ arguments, for any measurable function $\lambda : (\mathcal{X}\times\mathbb{R})^{2N} \to \mathbb{R}_+^*$ which is such that, for any $i \in \{1,...,2N\}$:*

$$\lambda(Z_1,...,Z_{2N}) = \lambda(Z_1,...,Z_{i-1},Z_{i+N},Z_{i+1},...,Z_{i+N-1},Z_i,Z_{i+N+1},...,Z_{2N}),$$

for any $\theta\in\Theta$:

$$\mathcal{P}\exp\left\{\frac{\mathcal{P}^{(2)}\lambda}{N}\sum_{i=1}^{N}\left[T_\theta(Z_i) - T_\theta(Z_{i+N})\right]\right.$$
$$\left. - \mathcal{P}^{(2)}\left[\frac{\lambda^2}{2N^2}\frac{1}{N}\sum_{i=1}^{N}\left[T_\theta(Z_i) - T_\theta(Z_{i+N})\right]^2\right] - \eta\right\} \leq \mathcal{P}\exp\left(-\eta\right)$$

and the reverse inequality.

Proof. Let $\mathcal{L}hs$ denote the left-hand side of lemma 5.3. For short, let us put:

$$s(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big[f_\theta(X_{i+N})Y_{i+N} - f_\theta(X_i)Y_i\Big]^2 = \frac{1}{N}\sum_{i=1}^{N}\left[T_\theta(Z_i) - T_\theta(Z_{i+N})\right]^2.$$

Then we have:

$$\begin{aligned}\mathcal{L}hs &= P_{2N}\exp P^{(2)}\left(\frac{\lambda}{N}\sum_{i=1}^{N}\Big[T_\theta(Z_i) - T_\theta(Z_{i+N})\Big] - \frac{\lambda^2}{2N}s(\theta) - \eta\right)\\ &\le P_{2N}P^{(2)}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\Big[T_\theta(Z_i) - T_\theta(Z_{i+N})\Big] - \frac{\lambda^2}{2N}s(\theta) - \eta\right),\end{aligned}$$

by Jensen's conditional inequality. Now, we can conclude as in lemma 5.1:

$$\begin{aligned}\mathcal{L}hs &= P_{2N}\exp\left(\sum_{i=1}^{N}\log\cosh\left\{\frac{\lambda}{N}\Big[T_\theta(Z_i) - T_\theta(Z_{i+N})\Big]\right\} - \frac{\lambda^2}{2N}s(\theta) - \eta\right)\\ &\le P_{2N}\exp\left(\frac{\lambda^2}{2N^2}\sum_{i=1}^{N}\Big[T_\theta(Z_i) - T_\theta(Z_{i+N})\Big]^2 - \frac{\lambda^2}{2N}s(\theta) - \eta\right)\\ &= P_{2N}\exp\Big(-\eta\Big).\end{aligned}$$

□

Proof of theorem 4.4. We apply both inequalities of lemma 5.3 to every $\theta_k, k \in \{1,...,m\}$, and we take:

$$\lambda = \sqrt{\frac{2N\log\frac{2m}{\varepsilon}}{s(\theta)}}.$$

We obtain, for any $k \in \{1,...,m\}$:

$$\mathcal{P}\exp\left\{\frac{\mathcal{P}^{(2)}\lambda}{N}\sum_{i=1}^{N}\left[T_\theta(Z_i) - T_\theta(Z_{i+N})\right] - \log\frac{2m}{\varepsilon} - \eta\right\} \le \varepsilon.$$

Or, with probability at least $1-\varepsilon$, for any k:

$$\frac{1}{N}\sum_{i=1}^{N}\left[T_\theta(Z_i) - T_\theta(Z_{i+N})\right] \le \sqrt{\frac{2\log\frac{2m}{\varepsilon}}{N}}\left[\mathcal{P}^{(2)}\left(s(\theta)^{-\frac{1}{2}}\right)\right]^{-1},$$

so:

$$\left[\frac{1}{N}\sum_{i=1}^{N}T_\theta(Z_i) - \frac{1}{N}\sum_{i=N+1}^{2N}T_\theta(Z_i)\right]^2 \le \frac{2\log\frac{2m}{\varepsilon}}{N}\mathcal{P}^{(2)}s(\theta).$$

We end the first part of the proof by noting that:

$$\mathcal{P}^{(2)}s(\theta) = V_1(\theta) + V_2(\theta) + \left[\frac{1}{N}\sum_{i=1}^{N}T_\theta(Z_i) - \frac{1}{N}\sum_{i=N+1}^{2N}T_\theta(Z_i)\right]^2.$$

Now, let us see how we can obtain the second part of the theorem. Note that:

$$V_2(\theta) = \frac{1}{N}\sum_{i=N+1}^{2N}T_\theta(Z_i)^2 - \left(\frac{1}{N}\sum_{i=N+1}^{2N}T_\theta(Z_i)\right)^2.$$

We upper bound the first term by using lemma 5.1 with $g(f_\theta(X_i), Y_i) = f_\theta(X_i)^2Y_i^2 = T_\theta(Z_i)^2$, so with probability at least $1-\varepsilon$, for any k:

$$\frac{1}{N}\sum_{i=N+1}^{2N}T_\theta(Z_i)^2 \le \frac{1}{N}\sum_{i=1}^{N}T_\theta(Z_i)^2 + \sqrt{\frac{2\log\frac{m}{\varepsilon}\frac{1}{N}\sum_{i=1}^{2N}T_\theta(Z_i)^4}{N}}.$$

For the second order term, we use both inequalities of lemma 5.1 with $g(f_\theta(X_i), Y_i) = f_\theta(X_i)Y_i = T_\theta(Z_i)$, so with probability at least $1-\varepsilon$, for any k::

$$\left(\frac{1}{N}\sum_{i=1}^{N} T_\theta(Z_i)\right)^2 - \left(\frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)\right)^2$$
$$\leq \left|\frac{1}{N}\sum_{i=1}^{N} T_\theta(Z_i) - \frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)\right| \left|\frac{1}{N}\sum_{i=1}^{2N} T_\theta(Z_i)\right|$$
$$\leq 2\sqrt{\frac{\frac{1}{N}\sum_{i=1}^{2N} T_\theta(Z_i)^2 \log\frac{2m}{\varepsilon}}{N}} \frac{1}{N}\sum_{i=1}^{2N} |T_\theta(Z_i)|.$$

Putting all pieces together (and replacing ε by $\varepsilon/3$) ends the proof. □

5.4. **Proof of theorem 4.5.**

Proof of theorem 4.5. We introduce the following conditional probability measures, for any $i \in \{1, ..., N\}$:

$$\mathbb{P}_i = \frac{1}{(k+1)!}\sum_{\sigma\in\mathfrak{S}_{k+1}} \delta_{(Z_1,...,Z_{i-1},Z_{N(\sigma(1)-1)+i},Z_{i+1},...,}$$
$$Z_{N+i-1},Z_{N(\sigma(2)-1)+i},Z_{N+i+1},...,...,Z_{kN+i-1},Z_{N(\sigma(k+1)-1)+i},Z_{kN+i+1},...,Z_{(k+1)N})}.$$

and:

$$\mathbb{P} = \bigotimes_{i=1}^{N} \mathbb{P}_i$$

and, finaly, remember that:

$$\mathbf{P} = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N} \delta_{Z_i}.$$

Note that, by exchangeability, for any nonnegative function

$$h : (\mathcal{X}\times\mathbb{R})^{(k+1)N} \to \mathbb{R}$$

we have, for any $i \in \{1, ..., N\}$:

$$P_{(k+1)N}\mathbb{P}_i h(Z_1, ..., Z_{2N}) = P_{(k+1)N} h(Z_1, ..., Z_{2N}).$$

Lemma 5.4. *Let χ be a function $\mathbb{R} \to \mathbb{R}$. For any exchangeable functions λ, $\eta : (\mathcal{X}\times\mathbb{R})^{(k+1)N} \to \mathbb{R}_+$ and $\theta : (\mathcal{X}\times\mathbb{R})^{(k+1)N} \to \Theta$ we have:*

$$\mathbb{P}\exp\left\{\lambda\left[\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\chi\Big[f_\theta(X_i)Y_i\Big] - \frac{1}{N}\sum_{i=1}^{N}\chi\Big[f_\theta(X_i)Y_i\Big]\right] - \eta\right\}$$
$$\leq \exp(-\eta)\exp\left\{\frac{\lambda^2(1+k)^2}{2Nk^2}\mathbf{P}\left\{\Big[\chi\Big(f_\theta(X)Y\Big) - \mathbf{P}\chi\Big(f_\theta(X)Y\Big)\Big]^2\right\}\right.$$
$$\left. + \frac{\lambda^3(1+k)^3}{6N^2k^3}\left[\sup_{i\in\{1,...,(k+1)N\}}\chi\left(f_\theta(X_i)Y_i\right) - \inf_{i\in\{1,...,(k+1)N\}}\chi\left(f_\theta(X_i)Y_i\right)\right]^3\right\},$$

where we put $\lambda = \lambda(Z_1, ..., Z_{(k+1)N})$, $\theta = \theta(Z_1, ..., Z_{(k+1)N})$ and $\eta = \eta(Z_1, ..., Z_{(k+1)N})$ for short. We have the reverse inequality as well.

Before giving the proof, let us introduce the following useful notations.

Definition 5.3. We put, for any $\theta \in \Theta$, for any function χ:

$$\chi_i^\theta = \chi\left(Y_i f_\theta(X_i)\right),$$

and:

$$\chi^\theta = \chi\left(Y f_\theta(X)\right)$$

that means that:

$$\mathbf{P}\chi^\theta = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\chi_i^\theta.$$

We also put:

$$\mathcal{S}_\chi(\theta) = \sup_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta - \inf_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta.$$

Proof of the lemma. Remark that, for any exchangeable functions λ, $\eta : (\mathcal{X}\times\mathbb{R})^{(k+1)N}\to\mathbb{R}_+$ and $\theta : (\mathcal{X}\times\mathbb{R})^{kN}\to\Theta$ we have:

$$\begin{aligned}
&\mathbb{P}\exp\left\{\lambda\left[\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} g\Big[f_\theta(X_i)Y_i\Big] - \frac{1}{N}\sum_{i=1}^{N} g\Big[f_\theta(X_i)Y_i\Big]\right] - \eta\right\}\\
&\qquad = \exp\left(-\eta\right)\prod_{i=1}^{N}\mathbb{P}_i\exp\left\{\frac{\lambda}{kN}\sum_{j=1}^{k}\chi_{i+jN}^\theta - \frac{\lambda}{N}\chi_i^\theta\right\}\\
&\qquad = \exp\left(-\eta\right)\prod_{i=1}^{N}\exp\left\{\frac{\lambda}{kN}\sum_{j=0}^{k}\chi_{i+jN}^\theta\right\}\prod_{i=1}^{N}\mathbb{P}_i\exp\left\{-\frac{\lambda(1+k)}{kN}\chi_i^\theta\right\}
\end{aligned}$$

where we put $\lambda = \lambda(Z_1,\dots,Z_{kN})$, $\theta = \theta(Z_1,\dots,Z_{kN})$ and $\eta = \eta(Z_1,\dots,Z_{kN})$ for short.

Now, we have:

$$\log\prod_{i=1}^{N}\mathbb{P}_i\exp\left\{-\frac{\lambda(1+k)}{kN}\chi_i^\theta\right\} = \sum_{i=1}^{N}\log\mathbb{P}_i\exp\left\{-\frac{\lambda(1+k)}{kN}\chi_i^\theta\right\},$$

and, for any $i\in\{1,\dots,N\}$:

$$\begin{aligned}
&\log\mathbb{P}_i\exp\left\{-\frac{\lambda(1+k)}{Nk}\chi_i^\theta\right\}\\
&\qquad = -\frac{\lambda(1+k)}{Nk}\mathbb{P}_i\chi_i^\theta + \frac{\lambda^2(1+k)^2}{2N^2k^2}\mathbb{P}_i\left[\left(\chi_i^\theta - \mathbb{P}_i\chi_i^\theta\right)^2\right]\\
&\qquad\quad - \int_0^{\frac{\lambda(1+k)}{Nk}}\frac{1}{2}\left(\frac{\lambda(1+k)}{Nk}-\beta\right)^2\frac{1}{\mathbb{P}_i\exp\left[-\beta\chi_i^\theta\right]}\\
&\qquad\qquad \mathbb{P}_i\left[\left(\chi_i^\theta - \frac{\mathbb{P}_i\left\{\chi_i^\theta\exp\left[-\beta\chi_i^\theta\right]\right\}}{\mathbb{P}_i\exp\left[-\beta\chi_i^\theta\right]}\right)^3\exp\left(-\beta\chi_i^\theta\right)\right]d\beta.
\end{aligned}$$

Note that, for any $\beta\geq 0$:

$$\begin{aligned}
&\frac{1}{\mathbb{P}_i\exp\left[-\beta\chi_i^\theta\right]}\mathbb{P}_i\left[\left(\chi_i^\theta - \frac{\mathbb{P}_i\left\{\chi_i^\theta\exp\left[-\beta\chi_i^\theta\right]\right\}}{\mathbb{P}_i\exp\left[-\beta\chi_i^\theta\right]}\right)^3\exp\left(-\beta\chi_i^\theta\right)\right]\\
&\qquad\qquad \leq \left[\sup_{j\in\{1,\dots,k\}}\chi_{i+(j-1)N}^\theta - \inf_{j\in\{1,\dots,k\}}\chi_{i+(j-1)N}^\theta\right]^3,
\end{aligned}$$

and so:

$$\log \prod_{i=1}^{N} \mathbb{P}_i \exp\left\{-\frac{\lambda(1+k)}{Nk}\chi_i^\theta\right\} \leq -\frac{1}{N}\sum_{i=1}^{N}\frac{\lambda(1+k)}{k}\mathbb{P}_i\chi_i^\theta + \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda^2(1+k)^2}{2Nk^2}\mathbb{P}_i\left[\left(\chi_i^\theta - \mathbb{P}_i\chi_i^\theta\right)^2\right] + \frac{\lambda^3(1+k)^3}{6N^2k^3}\left[\sup_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta - \inf_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta\right]^3.$$

Note that:

$$\mathbb{P}_i\chi_i^\theta = \frac{1}{k+1}\sum_{j=0}^{k}\chi_{i+jN}^\theta$$

and so:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i\chi_i^\theta = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\chi_i^\theta = \mathbf{P}\chi^\theta;$$

remark also that:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i\left[\left(\chi_i^\theta - \mathbb{P}_i\chi_i^\theta\right)^2\right] \leq \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\left[\chi_i^\theta - \left(\frac{1}{(k+1)N}\sum_{j=1}^{(k+1)N}\chi_j^\theta\right)\right]^2 = \mathbf{P}\left[\left(\chi^\theta - \mathbf{P}\chi^\theta\right)^2\right],$$

we obtain:

$$\mathbb{P}\exp\left\{\lambda\left[\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}f_\theta(X_i)Y_i - \frac{1}{N}\sum_{i=1}^{N}f_\theta(X_i)Y_i\right] - \eta\right\} = \exp(-\eta)\exp\left\{\frac{\lambda^2(1+k)^2}{2Nk^2}\mathbf{P}\left[\left(\chi^\theta - \mathbf{P}\chi^\theta\right)^2\right] + \frac{\lambda^3(1+k)^3}{6N^2k^3}\left[\sup_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta - \inf_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta\right]^3\right\}.$$

The proof of the reverse inequality is exactly the same. □

Let us choose here again χ such that $\chi(u) = u$, namely: $\chi = id$. By the use of a union bound argument on elements of Θ_0 we obtain, for any $\varepsilon > 0$, for any exchangeable function $\lambda : (\mathcal{X}\times\mathbb{R})^{(k+1)N} \to \mathbb{R}_+$, with probability at least $1-\varepsilon$, for any $h \in \{1,...,m\}$:

$$\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}f_{\theta_h}(X_i)Y_i - \frac{1}{N}\sum_{i=1}^{N}f_{\theta_h}(X_i)Y_i \leq \frac{\lambda\left(1+\frac{1}{k}\right)^2}{2N}\mathbf{P}\left[\left(\chi^{\theta_h} - \mathbf{P}\chi^{\theta_h}\right)^2\right] + \frac{\lambda^2\left(1+\frac{1}{k}\right)^3}{6N^2}\mathcal{S}_{id}(\theta_h)^3 + \frac{\log\frac{m}{\varepsilon}}{\lambda}.$$

Let us choose, for any $h \in \{1,...,m\}$:

$$\lambda = \sqrt{\frac{2N\log\frac{m}{\varepsilon}}{\left(1+\frac{1}{k}\right)^2\mathbf{P}\left[\left(\chi^{\theta_h} - \mathbf{P}\chi^{\theta_h}\right)^2\right]}},$$

the bound becomes:

$$\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)Y_i - \frac{1}{N}\sum_{i=1}^{N} f_{\theta_h}(X_i)Y_i \leq \left(1+\frac{1}{k}\right)\left[2\sqrt{\frac{\mathbf{P}\left[\left(\chi^{\theta_h}-\mathbf{P}\chi^{\theta_h}\right)^2\right]\log\frac{m}{\varepsilon}}{2N}} + \frac{\mathcal{S}_{id}(\theta_h)^3\log\frac{m}{\varepsilon}}{3N\mathbf{P}\left[\left(\chi^{\theta_h}-\mathbf{P}\chi^{\theta_h}\right)^2\right]}\right].$$

We use the reverse inequality exactly in the same way, we then combine both inequality by a union bound argument and obtain the following result. For any $\varepsilon>0$, with $P_{(k+1)N}$ probability at least $1-\varepsilon$ we have, for any $h\in\{1,...,m\}$:

$$r_2\left(\mathcal{C}^h\alpha_1^h\theta_h\right) - r_2\left(\alpha_2^h\theta_h\right) \leq \frac{\left(1+\frac{1}{k}\right)^2}{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)^2}\left[\frac{2\mathbb{V}_{\theta_h}\log\frac{2m}{\varepsilon}}{N} + \frac{2\left(\log\frac{2m}{\varepsilon}\right)^{\frac{3}{2}}\mathcal{S}_{id}(\theta_h)^3}{3N^{\frac{3}{2}}\mathbb{V}_{\theta_h}^{\frac{1}{2}}} + \frac{\left(\log\frac{2m}{\varepsilon}\right)^2\mathcal{S}_{id}(\theta_h)^6}{9N^2\mathbb{V}_{\theta_h}^2}\right], \tag{5.3}$$

remember that:

$$\mathbb{V}_\theta = \mathbf{P}\left\{\left[\Big(f_\theta(X)Y\Big) - \mathbf{P}\Big(f_\theta(X)Y\Big)\right]^2\right\}.$$

We now give a new lemma.

Lemma 5.5. *Let us assume that P is such that, for any $h\in\{1,...,m\}$:*

$$\exists\beta_h>0, \exists B_h\geq 0, P\exp\left(\beta_h\left|f_{\theta_h}(X)Y\right|\right)\leq B_h.$$

This if for example the case if $f_{\theta_h}(X_i)Y_i$ is subgaussian, with any $\beta_h>0$ and

$$B_h = 2\exp\left\{\frac{\beta_h^2}{2}P\left[\left(f_{\theta_h}(X)Y\right)^2\right]\right\}.$$

Then we have, for any $\varepsilon\geq 0$:

$$P_{(k+1)N}\left\{\sup_{1\leq i\leq(k+1)N} f_{\theta_h}(X_i)Y_i \leq \frac{1}{\beta_h}\log\frac{(k+1)NB_h}{\varepsilon}\right\}\geq 1-\varepsilon.$$

Proof of the lemma. We have:

$$\begin{aligned} &P_{(k+1)N}\left(\sup_{1\leq i\leq(k+1)N} f_{\theta_h}(X_i)Y_i\geq s\right)\\ &\quad= P_{(k+1)N}\left(\exists i\in\{1,...,(k+1)N\}, f_{\theta_h}(X_i)Y_i\geq s\right)\\ &\quad= \sum_{i=1}^{(k+1)N} P\mathbb{1}_{f_{\theta_h}(X_i)Y_i\geq s}\\ &\quad\leq (k+1)NP\exp\left(\beta_h\left|f_{\theta_h}(X_i)Y_i-s\right|\right)\leq (k+1)NB_h\exp\left(-\beta_h s\right). \end{aligned}$$

Now, let use choose:

$$s=\frac{1}{\beta_h}\log\frac{(k+1)NB_h}{\varepsilon},$$

and we obtain the lemma. □

As a consequence, using a union bound argument, we have, for any $\varepsilon \geq 0$, with probability at least $1-\varepsilon$, for any $h \in \{1, ..., m\}$:

$$\mathcal{S}_{id}(\theta_h) = \sup_{i \in \{1,...,(k+1)N\}} f_{\theta_h}(X_i)Y_i - \inf_{i \in \{1,...,(k+1)N\}} f_{\theta_h}(X_i)Y_i \leq \frac{2}{\beta_h} \log \frac{2(k+1)mNB_h}{\varepsilon}.$$

By pluggin the lemma into equation 5.3 we obtain the theorem. □

6. Simulations in the transductive case

6.1. **Description of the example.** Here, we assume that we have:

$$Y_i = f(X_i) + \xi_i$$

for $i \in \{1, ..., 2N\}$ with $N = 2^{10} = 1024$, where the variables $X_i \in [0,1] \subset \mathbb{R}$ are i.i.d. from a uniform distribution $\mathcal{U}(0,1)$ (here we DO NOT assume that the statistician knows this point), the η_i are i.i.d. from a gaussian distribution $\mathcal{N}(0,\sigma)$ and independant from the X_i. The statistician observes $(X_1, Y_1), ..., (X_N, Y_N)$ and $X_{N+1}, ..., X_{2N}$ and wants to estimate $Y_{N+1}, ..., Y_{2N}$.

We will here again use three estimations methods: an inductive method, that does not take advantage of the knowledge of $X_{N+1}, ..., X_{2N}$, and two transductive methods. For the inductive method, we take the tresholded wavelet estimator that we used in the experiments in the inductive case. For the transductive method, we use here again a wavelet estimator and a (multiscale) SVM.

6.2. **The estimators.**

6.2.1. *Thresholded wavelets estimators.* In this case, as we assume that we don't know the distribution $P_{(X)}$, we have to estimate it and use a warped wavelet estimator. We take:

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i \leq x),$$

and:

$$\hat{\beta}_{j,k} = \frac{1}{N} \sum_{j=1}^{N} Y_i \psi_{j,k}(F_N(X_i)),$$

$$\tilde{f}_J(.) = \sum_{j=-1}^{J} \sum_{k \in S_j} \hat{\beta}_{j,k} \mathbb{1}(|\hat{\beta}_{j,k}| \geq \kappa t_N) \psi_{j,k}(F_N(.)).$$

Here again, we choose $\kappa = 0.5$ and $J = 7$.

6.2.2. *Wavelet estimators with our algorithm.* Here, we use the same family of functions, and we apply the transductive method described previously. Here again, we use gaussian approximations for the confidence intervals (but we double their length in order to take into account the variance of both samples).

6.2.3. *SVM estimator.* The transductive SVM estimator is taken with kernel:

$$K_\gamma(x, x') = \exp\left(-2^{2\gamma}(x - x')^2\right)$$

and $\gamma \in \{1, ..., m'(N)\}$ where $m'(N) = 6$. We use the same gaussian approximation than in the previous example.

6.3. **Experiments and results.** We consider the same functions than in the inductive case. We choose ε=10%. We repeat each experiment 20 times. We give the results in figure 8.

FIGURE 8. Results of the experiments. For each experiment, we give the mean risk r_2.

Function $f(.)$	s.d. σ	"inductive" thresholded wavelets	transductive thresh. wav. with our method	transductive multiscale SVM
Doppler	0.3	0.234	0.174	0.165
HeaviSine	0.3	0.134	0.156	0.134
Blocks	0.3	0.187	0.171	0.177
Doppler	1	1.179	1.152	1.120
HeaviSine	1	1.092	1.110	1.065
Blocks	1	1.153	1.144	1.129

7. Bound on a multidimensionnal model

7.1. **Theorem and algorithm.** In this subsection, we try to generalize the algorithm described in section 4 to the case where there are multidimensional models. The idea is that, for example, if $\Theta_0 = \{\theta_1, \theta_2, \theta_3\}$, we could try not only to make projections on:

$$\{\alpha\theta_i, \alpha \in \mathbb{R}\} \text{ for } i \in \{1, 2, 3\}$$

but also on a bidimensional space like:

$$\left\{\alpha\theta_1 + \beta\theta_2, (\alpha, \beta) \in \mathbb{R}^2\right\}.$$

More precisely, let us give the following definitions. First of all, we assume that we are in the case where $k = 1$, so the test sample and the learning sample have size N. We always assume that:

$$\Theta_0(Z_1, ..., Z_{2N}) = \{\theta_1, ..., \theta_m\}$$

is such that every θ_k is an exchangeable function of $(Z_1, ..., Z_{2N})$.

Definition 7.1. For every $d \geq 0$, $0 < j_1 < ... < j_d < m+1$ and $\mathcal{S} = (\theta_{j_1}, ..., \theta_{j_d}) \in \Theta_0^d$ we put:

$$f_{\mathcal{S}}(x) = \left(f_{\theta_{j_1}}(x), ..., f_{\theta_{j_d}}(x)\right).$$

For convenience, let us put, for any $\alpha = (\alpha^1, ..., \alpha^d) \in \mathbb{R}^d$:

$$\alpha\mathcal{S}' = \sum_{k=1}^{d} \alpha^k \theta_{j_k}$$

Remark that we have:

$$\alpha f_{\mathcal{S}}(.)' = f_{\alpha\mathcal{S}'}(.) : \mathcal{X} \to \mathbb{R};$$

let us put:

$$C_{1,2}^{\mathcal{S}} = \frac{1}{N} \sum_{i=1}^{2N} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)$$

$$C_1^{\mathcal{S}} = \frac{1}{N} \sum_{i=1}^{N} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)$$

$$\mathcal{M}^{\mathcal{S}} = \frac{1}{2} C_{1,2}^{\mathcal{S}} \left(C_1^{\mathcal{S}}\right)^{-1},$$

and finally:

$$\alpha_{1,2}^{\mathcal{S}} = \arg\min_{\alpha\in\mathbb{R}^d} r_{1,2}\left(\alpha\mathcal{S}'\right) = \frac{1}{N}\sum_{i=1}^{2N} Y_i f_{\mathcal{S}}(X_i)\left(C_{1,2}^{\mathcal{S}}\right)^{-1}$$

$$\alpha_{1}^{\mathcal{S}} = \arg\min_{\alpha\in\mathbb{R}^d} r_{1}\left(\alpha\mathcal{S}'\right) = \frac{1}{N}\sum_{i=1}^{N} Y_i f_{\mathcal{S}}(X_i)\left(C_{1}^{\mathcal{S}}\right)^{-1}.$$

For any matrix M we will let $\rho(M)$ denote the biggest eigenvalue of M.

Here, $\alpha_{1,2}^{\mathcal{S}}$ is our objective but we can only observe $\alpha_1^{\mathcal{S}}$, and the matrix $\mathcal{M}^{\mathcal{S}}$.

Remark 7.1. Note the change in the objective. In this subsection, we try to minimize $r_{1,2}$ and not r_2.

Theorem 7.1. *Let $d \geq 0$, let $\mathcal{S} \in \Theta^d$. Let us put:*

$$B_{1,2}^{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{2N} Y_i^2 \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)\mathcal{C}_{1,2}^{-\frac{1}{2}}.$$

For any $\varepsilon > 0$, we have, with P_{2N}-probability at least $1-\varepsilon$:

$$r_{1,2}\left(\mathcal{M}^{\mathcal{S}}\alpha_1^{\mathcal{S}}\mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}}\mathcal{S}'\right) \leq \frac{4\rho\left(B_{1,2}^{\mathcal{S}}\right)}{N}\left(d\log(2) + 2\log\frac{1}{\varepsilon}\right).$$

Note that $B_{1,2}^{\mathcal{S}}$ is not observable, except in the case of classification where we have $Y_i \in \{-1,+1\}$ and so $Y_i^2 = 1$, which implies that $B_{1,2}^{\mathcal{S}} = I$ and so:

$$\rho\left(B_{1,2}^{\mathcal{S}}\right) = 1.$$

In the general case we have the following corollary.

Corollary 7.2. *Let $d \geq 0$, let $\mathcal{S} \in \Theta^d$. Let us put:*

$$B_{1}^{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)\mathcal{C}_{1,2}^{-\frac{1}{2}},$$

that is observable, and:

$$D_{1,2}^{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{2N} Y_i^4\left(\lambda_{1,2}C_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i) C_{1,2}^{-\frac{1}{2}}\lambda_{1,2}'\right)^2,$$

where:

$$\rho\left(B_{1,2}\right) = \sup_{\|\lambda\|=1} \lambda B_{1,2}\lambda' = \lambda_{1,2}B_{1,2}\lambda_{1,2}'.$$

For any $\varepsilon > 0$, we have with P_{2N}-probability at least $1-\varepsilon$:

$$r_{1,2}\left(\mathcal{M}^{\mathcal{S}}\alpha_1^{\mathcal{S}}\mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}}\mathcal{S}'\right) \leq \frac{8\rho\left(B_1^{\mathcal{S}}\right)}{N}\left(d\log(2) + 2\log\frac{2}{\varepsilon}\right) + \frac{4\left[D_{1,2}^{\mathcal{S}} + \log\frac{2}{\varepsilon}\right]}{N^{\frac{3}{2}}}.$$

We can now give a new algorithm to perform regression estimation, that is a variant of the one given in section 4. Before all, we have to choose k dimensions $d_1, ..., d_k$ and k models

$$\mathcal{S}_1 \in \Theta^{d_1}, ..., \mathcal{S}_k \in \Theta^{d_k}.$$

We apply theorem 7.1 to all the models simultaneously by a union bound argument and we obtain k confidence regions:

$$\mathcal{CR}_1, ..., \mathcal{CR}_k$$

and the corresponding projections:

$$\Pi_1, ..., \Pi_k.$$

We then use the following algorithm:

- choose $\theta(0) = 0$;
- at step $n \in \mathbb{N}^*$, define:

$$k'(n) = \arg\max_{k'} d_{1,2}\left(\theta(n-1), \Pi_{k'}\theta(n-1)\right)$$

and

$$\theta(n) = \Pi_{k'(n)}\theta(n-1);$$

- stop when $d_{1,2}\left(\theta(n), \theta(n-1)\right) \leq \kappa$.

Example 7.1. By taking $k = m$, $d_1 = ... = d_m = 1$ and $\mathcal{S}_i = \theta_i$ for all i, we obtain exactly the projection algorithm described in section 4.

Example 7.2. Let us take $k = m$, $d_i = i$ for any i and $\mathcal{S}_i = (\theta_1, ..., \theta_i)$: we are in the case of nested submodels, and we obtain a procedure similar to Lepski's method, at least as it is described by Birg [3].

7.2. **Proofs.** For convenience, we assume that $\mathcal{S}$ is chosen once and for all, and so we will let $B_{1,2}$ stand for $B_{1,2}^{\mathcal{S}}$, $\mathcal{C}_{1,2}$ for $\mathcal{C}_{1,2}^{\mathcal{S}}$, and $\mathcal{D}_{1,2}$ for $\mathcal{D}_{1,2}^{\mathcal{S}}$. We keep the notation $f_{\mathcal{S}}(.)$ to avoid confusion with the true regression function $f(.)$.

Proof of theorem 7.1. Let us state the following variant of lemma 5.1, obtained exactly in the same way. For any measurable function $\eta : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any measurable function $\gamma : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\lambda \in \mathbb{R}^d$:

$$P_{2N} \exp\left(\gamma \left\langle \frac{\mathcal{C}_{1,2}^{-\frac{1}{2}}}{N} \sum_{i=1}^{N} \Big\{ f_{\mathcal{S}}(X_i)Y_i - f_{\mathcal{S}}(X_{i+N})Y_{i+N} \Big\}, \lambda \right\rangle - \|\lambda\|^2 - \eta \right)$$
$$\leq P_{2N} \exp\left(\frac{\gamma^2}{N} \frac{1}{N} \sum_{i=1}^{2N} \left\langle \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)Y_i, \lambda \right\rangle^2 - \|\lambda\|^2 - \eta \right),$$

that can be written:

$$P_{2N} \exp\left(\gamma A\lambda' - \lambda I\lambda' - \eta\right) \leq P_{2N} \exp\left(\frac{\gamma^2}{N} \lambda B_{1,2}\lambda' - \lambda I\lambda' - \eta\right)$$

where:

$$A = \frac{\mathcal{C}_{1,2}^{-\frac{1}{2}}}{N} \sum_{i=1}^{N} \Big\{ \Psi(X_i)Y_i - \Psi(X_{i+N})Y_{i+N} \Big\}.$$

So we have:

$$\int_{\mathbb{R}^d} P_{2N} \exp\left(\gamma A\lambda' - \lambda I\lambda' - \eta\right) d\lambda \leq \int_{\mathbb{R}^d} P_{2N} \exp\left(\frac{\gamma^2}{N} \lambda B_{1,2}\lambda' - \lambda I\lambda' - \eta\right) d\lambda.$$

Using Fubini's theorem we obtain:

$$P_{2N} \int_{\mathbb{R}^d} \exp\left(\gamma A\lambda' - \lambda I\lambda' - \eta\right) d\lambda \leq P_{2N} \int_{\mathbb{R}^d} \exp\left(\lambda \left(\frac{\gamma^2}{N} B_{1,2} - I\right) \lambda' - \eta\right) d\lambda.$$

Now, let us assume that γ is small enough for the matrix

$$I - \frac{\gamma^2}{N} B_{1,2}$$

to be definite positive. Actually this means that:

$$\frac{N}{\gamma^2} > \rho\left(B_{1,2}\right)$$

or:

$$\gamma < \sqrt{\frac{N}{\rho\left(B_{1,2}\right)}}.$$

Then we get:

$$P_{2N}\left[\pi^{\frac{d}{2}}\exp\left(\frac{\gamma^2}{4}AA' - \eta\right)\right] \leq P_{2N}\left[\frac{\pi^{\frac{d}{2}}\exp(-\eta)}{\sqrt{\det\left(I - \frac{\gamma^2}{N}B_{1,2}\right)}}\right],$$

or:

$$P_{2N}\left[\exp\left(\frac{\gamma^2}{4}AA' - \eta\right)\right] \leq P_{2N}\left[\exp\left(-\eta - \frac{1}{2}\log\det\left(I - \frac{\gamma^2}{N}B_{1,2}\right)\right)\right].$$

Let us put:

$$\eta = -\frac{1}{2}\log\det\left(I - \frac{\gamma^2}{N}B_{1,2}\right) + \log\frac{1}{\varepsilon},$$

we get:

$$P_{2N}\left[\exp\left(\frac{\gamma^2}{4}AA' + \frac{1}{2}\log\det\left(I - \frac{\gamma^2}{N}B_{1,2}\right) - \log\frac{1}{\varepsilon}\right)\right] \leq \varepsilon.$$

This implies that:

$$P_{2N}\left[AA' \leq \frac{4}{\gamma^2}\log\frac{1}{\varepsilon\sqrt{\det\left(I - \frac{\gamma^2}{N}B\right)}}\right] \geq 1 - \varepsilon.$$

Finally, note that:

$$AA' = r_{1,2}\left(\mathcal{M}^{\mathcal{S}}\alpha_1^{\mathcal{S}}\mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}}\mathcal{S}'\right).$$

We obtain the following result. For any $\varepsilon > 0$, for any measurable function $\gamma : (\mathcal{X}\times\mathbb{R})^{2N} \to \mathbb{R}_+$ that is exchangeable with respect to its $2\times 2N$ arguments:

$$P_{2N}\left[r_{1,2}\left(\mathcal{M}^{\mathcal{S}}\alpha_1^{\mathcal{S}}\mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}}\mathcal{S}'\right) \leq \frac{4}{\gamma^2}\log\frac{1}{\varepsilon\sqrt{\det\left(I - \frac{\gamma^2}{N}B_{1,2}\right)}}\right] \geq 1 - \varepsilon.$$

In particular if we choose:

$$\gamma = \sqrt{\frac{N}{2\rho\left(B_{1,2}\right)}}$$

then we obtain the theorem. □

Proof of corollary 7.2. In a first time, let us introduce the following obvious notation:

$$B_2^{\mathcal{S}} = B_2 = \frac{1}{N}\sum_{i=N+1}^{2N} Y_i^2 \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)\mathcal{C}_{1,2}^{-\frac{1}{2}}.$$

Now, we state a new variant of lemma 5.1: For any measurable function $\eta : (\mathcal{X}\times\mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2\times 2N$ arguments, for any

measurable function $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}^d$ that is exchangeable with respect to its $2 \times 2N$ arguments:

$$P_{2N} \exp\left(\lambda B_2 \lambda' - \lambda B_1 \lambda' - \eta \right) \leq P_{2N} \exp\left(\frac{1}{N^2} \sum_{i=1}^{2N} Y_i^4 \left(\lambda C_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i) C_{1,2}^{-\frac{1}{2}} \lambda' \right)^2 - \eta \right).$$

Now, taking:

$$\lambda = N^{\frac{1}{4}} \lambda_{1,2}$$

and:

$$\eta = \log \frac{1}{\varepsilon}$$

we obtain:

$$P_{2N} \left[\lambda_{1,2} B_2 \lambda'_{1,2} \leq \lambda_{1,2} B_1 \lambda'_{1,2} + \frac{D_{1,2} + \log \frac{1}{\varepsilon}}{\sqrt{N}} \right] \geq 1 - \varepsilon.$$

So, with probability at least $1 - \varepsilon$:

$$\begin{aligned} \rho\left(B_{1,2}\right) = \lambda_{1,2} B_{1,2} \lambda'_{1,2} &= \lambda_{1,2} B_1 \lambda'_{1,2} + \lambda_{1,2} B_2 \lambda'_{1,2} \\ &\leq 2 \lambda_{1,2} B_1 \lambda'_{1,2} + \frac{D_{1,2} + \log \frac{1}{\varepsilon}}{\sqrt{N}} \\ &\leq 2 \sup_{\|\lambda\|=1} \lambda B_1 \lambda' + \frac{D_{1,2} + \log \frac{1}{\varepsilon}}{\sqrt{N}} = 2\rho\left(B_1\right) + \frac{D_{1,2} + \log \frac{1}{\varepsilon}}{\sqrt{N}}. \end{aligned}$$

The last step is to combine this inequality with theorem 7.1 by a union bound argument. □

8. Interpretation of theorem 2.4 as an oracle inequality

We conclude this paper by going back to the inductive case. We first give a weak variant of theorem 2.1, in order to obtain an easily observable bound. We then use theorem 2.4 as an oracle inequality to show that the obtained estimator is adaptative, which means that if we assume that the true regression function f has an unknown regularity β, then the estimator is able to reach the right speed of convergence $N^{\frac{-2\beta}{2\beta+1}}$ up to a $\log N$ factor.

8.1. A weak version of theorem 2.1.

Let us assume that $\mathcal{X} = [0,1]$ and let us put $\Theta = \mathbb{L}_2(P_{(X)})$. Let $(\theta_k)_{k \in \mathbb{N}^*}$ be an orthonormal basis of Θ, and we simply take, for any x and θ:

$$f_\theta(x) = \theta(x),$$

that m is chosen and we still have:

$$\Theta_0 = (\theta_1, ..., \theta_m).$$

Moreover, let us assume that P is such that $Y_i = f(X_i) + \eta_i$ where η_i is independant of X_i and has an unknown distribution, with of course $P\eta = 0$ and $P(\eta^2) =\leq \sigma^2 < \infty$ with a known σ. We do not assume stronger hypothesis about η.

Theorem 8.1. *We have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $k \in \{1, ..., m\}$:*

$$R(\mathcal{C}_k \hat{\alpha}_k \theta_k) - R(\overline{\alpha}_k \theta_k) \leq \frac{4\left[1 + \log \frac{2m}{\varepsilon}\right]}{N} \left[\frac{1}{N} \sum_{i=1}^{N} f_{\theta_k}(X_i)^2 Y_i^2 + B^2 + \sigma^2 \right].$$

The proof is given at the end of the section. Note that this theorem is more general than theorem 2.1 in the following way: we do not require the existence of exponential moments for the noise η_i. But, at least for large values of N, the bound is less tight.

8.2. **Rate of convergence of the obtained estimator.** Now, let us put:

$$\overline{\theta}_m = \arg \min_{\theta \in Span(\Theta_0)} R(\theta)$$

(that depends effectively on m by $\Theta_0 = \{\theta_1, ..., \theta_m\}$), and let us assume that f satisfies the two following conditions: it is regular, namely there is an unknown $\beta \geq 1$ and a $C \geq 0$ such that:

$$\left\| f_{\overline{\theta}_m} - f \right\|_P^2 \leq C m^{-2\beta},$$

and that we have a constant $B < \infty$ such that:

$$\sup_{x \in \mathcal{X}} f(x) \leq B$$

with B known to the statistician. It follows that:

$$\|f\|_P^2 \leq B^2.$$

If follows that every set, for $k \in \{1, ..., m\}$:

$$\mathcal{F}_k = \left\{ \sum_{j=1}^{\infty} \alpha_j \theta_j : \alpha_k^2 \leq B^2 \right\} \bigcap \Theta$$

is a convex set that contains f and so that the orthogonal projection: $\Pi_P^{\mathcal{F},m} = \Pi_P^{\mathcal{F}_m} ... \Pi_P^{\mathcal{F}_1}$ (where $\Pi_P^{\mathcal{F}_k}$ denotes the orthogonal projection on $\mathcal{F}_k$) can only improve an estimator:

$$\forall \theta, \left\| \Pi_P^{\mathcal{F},m} \theta - f \right\|_P^2 \leq \|\theta - f\|_P^2 .$$

Actually, note that this projection just consists in thresholding very large coefficients to a limited value. This modification is necessary in what follows, but this is just a technical remark: most of the time, our estimator won't be modified by $\Pi_P^{\mathcal{F},m}$ for any m.

Remember also that in this context, the estimator given in definition 2.5 is just:

$$\hat{f}(x) = f_{\hat{\theta}}(x),$$

with:

$$\hat{\theta} = \Pi_P^{m,\varepsilon} ... \Pi_P^{1,\varepsilon} 0.$$

Theorem 8.2. *Let us assume that $\Theta = \mathbb{L}_2(P_{(X)})$, $\mathcal{X} = [0,1]$ and $(\theta_k)_{k \in \mathbb{N}^*}$ is an orthonormal basis of Θ. Let us assume that we are in the idealized regression model:*

$$Y = f(X) + \eta,$$

where $P\eta = 0$, $P(\eta^2) \leq \sigma^2 < \infty$ and η and X are independant, and σ is known. Let us assume that $f \in \Theta$ is such that there is an unknown $\beta \geq 1$ and an unknown $C \geq 0$ such that:

$$\left\| f_{\overline{\theta}_m} - f \right\|_P^2 \leq C m^{-2\beta},$$

and that we have a constant $B < \infty$ such that:

$$\sup_{x \in \mathcal{X}} f(x) \leq B$$

with B known to the statistician. Then our estimator $\hat{f}$ (given in definition 2.5 with $n_0 = m$ here, build using the bound $\beta(\varepsilon, k)$ given in theorem 8.1), with $\varepsilon = N^{-2}$ and $m = N$, is such that, for any $N \geq 2$,

$$P^{\otimes N}\left[\left\|\Pi_P^{\mathcal{F},N}\hat{f} - f\right\|_P^2\right] \leq C'(C,B,\sigma)N^{\frac{-2\beta}{2\beta+1}}\log N,$$

where we have:

$$C'(C,B,\sigma) = 2C + 50B^2 + 44\sigma^2.$$

Remark that this theorem shows that the estimator is able to achieve the good rate of convergence, up to a $\log N$ factor, with an unknown β.

We can remark too that the given $C'(C,B,\sigma)$ isn't the best possible. It is easy to improve C', but this is pointless here. The reason is that theorem 8.2 has only an asymptotic interest. In order to have good results for a given N, we have to use theorem 2.2 that gives numerically better results than its weaker version, theorem 8.1. The reason why we use the weaker version here is that it is more easy in this context to study asymptotics properties.

8.3. **Proof of the theorems: theorem 2.4 used as an oracle inequality.**

Proof of theorem 8.1. Actually, the proof is quite straightforward: instead of using the techniques given in the section devoted to the inductive case, we use a result valid in the transductive case and integrate it with respect to the test sample. There are several ways to perform this integration (see for example Catoni [5]), here we choose to apply a result obtained by Panchenko [10] that gives a particularly simple result here.

Lemma 8.3 (Panchenko [10], corollary 1)**.** *Let us assume that we have i.i.d. variables $T_1, ..., T_N$ (with distribution P and values in $\mathbb{R}$) and an independant copy $T' = (T'_1, ..., T'_N)$ of $T = (T_1, ..., T_N)$. Let $\xi_j(T, T')$ for $j \in \{1, 2, 3\}$ be three measurables functions taking values in $\mathbb{R}$, and $\xi_3 \geq 0$. Let us assume that we know two constants $A \geq 1$ and $a > 0$ such that, for any $u > 0$:*

$$P^{\otimes 2N}\left[\xi_1(T,T') \geq \xi_2(T,T') + \sqrt{\xi_3(T,T')u}\right] \leq A\exp(-au).$$

Then, for any $u > 0$:

$$P^{\otimes 2N}\Bigg\{P^{\otimes 2N}\left[\xi_1(T,T')|T\right] \\ \geq P^{\otimes 2N}\left[\xi_2(T,T')|T\right] + \sqrt{P^{\otimes 2N}\left[\xi_3(T,T')|T\right]u}\Bigg\} \leq A\exp(1-au).$$

Now, a simple application of the first inequality of lemma 5.1 (given in the transductive section) with $\varepsilon > 0$, any $k \in \{1, ..., m\}$, $g = id$, $\eta = 1 + \log\frac{2m}{\varepsilon}$ and:

$$\lambda_k = \sqrt{\frac{N\eta}{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2}}$$

leads us to the following bound, for any k:

$$P^{\otimes 2N}\exp\left[\sqrt{N\eta}\frac{\frac{1}{N}\sum_{i=1}^{N}\left[f_{\theta_k}(X_i)Y_i - f_{\theta_k}(X_{i+N})Y_{i+N}\right]}{\sqrt{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2}} - 2\eta\right] \leq \exp(-\eta),$$

or:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}[f_{\theta_k}(X_i)Y_i - f_{\theta_k}(X_{i+N})Y_{i+N}] \geq \sqrt{\frac{4\eta}{N^2}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2}\right] \leq \exp(-\eta) = \frac{\varepsilon}{2k\exp(1)}.$$

We now apply Panchenko lemma with:

$$T_i = f_{\theta_k}(X_i)Y_i, \quad T_i' = f_{\theta_k}(X_{i+N})Y_{i+N}$$

$$\xi_1(T,T') = \frac{1}{N}\sum_{i=1}^{N} T_i, \quad \xi_2(T,T') = \frac{1}{N}\sum_{i=1}^{N} T_i',$$

$$\xi_3(T,T') = \frac{2}{N^2}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^2 Y_i^2 \geq 0,$$

and $A = a = 1$. We obtain:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}[f_{\theta_k}(X_i)Y_i - P[f_{\theta_k}(X)Y]] \geq \sqrt{\frac{4\eta}{N^2}\sum_{i=1}^{N}[f_{\theta_k}(X_i)^2Y_i^2 + P[f_{\theta_k}(X)^2Y^2]]}\right] \leq \exp(1-\eta) = \frac{\varepsilon}{2k}.$$

Remark finally that:

$$P\left[f_{\theta_k}(X)^2Y^2\right] \leq P\left[f_{\theta_k}(X)^2\right](B^2+\sigma^2),$$

and by the orthonormality property of the basis $(\theta_k)_{k\geq 1}$:

$$P\left[f_{\theta_k}(X)^2\right] = 1.$$

We proceed exactly in the same way with the reverse inequalities for any k and combine the obtained $2m$ inequalities to obtain the result:

$$\begin{aligned} P^{\otimes N}\Bigg\{\exists k \in \{1,...,m\}, \frac{1}{N}\sum_{i=1}^{N}\Big|f_{\theta_k}(X_i)Y_i - P[f_{\theta_k}(X)Y]\Big| \\ \geq \sqrt{\frac{4+4\log\frac{2m}{\varepsilon}}{N^2}\sum_{i=1}^{N}[f_{\theta_k}(X_i)^2Y_i^2 + B^2 + \sigma^2]}\Bigg\} \\ = P^{\otimes 2N}\Bigg\{\exists k \in \{1,...,m\}, \frac{1}{N}\sum_{i=1}^{N}\Big|f_{\theta_k}(X_i)Y_i - P[f_{\theta_k}(X)Y]\Big| \\ \geq \sqrt{\frac{4+4\log\frac{2m}{\varepsilon}}{N^2}\sum_{i=1}^{N}[f_{\theta_k}(X_i)^2Y_i^2 + B^2 + \sigma^2]}\Bigg\} \leq \varepsilon \end{aligned}$$

that ends the proof. □

Proof of theorem 8.2. Let us begin the proof with a general m and ε, the reason of the choice $m = N$ and $\varepsilon = N^{-2}$ will become clear. Let us also call $\mathcal{E}(\varepsilon)$ the event

satisfied with probability at least $1-\varepsilon$ in theorem 8.1. We have:

$$P^{\otimes N}\left[\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2\right]=P^{\otimes N}\left[\mathbb{1}_{\mathcal{E}(\varepsilon)}\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2\right]$$
$$+P^{\otimes N}\left[\left(1-\mathbb{1}_{\mathcal{E}(\varepsilon)}\right)\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2\right].$$

First of all, it is obvious that:

$$P^{\otimes N}\left[\left(1-\mathbb{1}_{\mathcal{E}(\varepsilon)}\right)\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2\right]$$
$$\leq 2P^{\otimes N}\left[\left(1-\mathbb{1}_{\mathcal{E}(\varepsilon)}\right)\left(\left\|\Pi_P^{\mathcal{F},m}\hat{f}\right\|_P^2+\|f\|_P^2\right)\right]$$
$$\leq 2\varepsilon\left(B^2m+B^2\right)=2\varepsilon(m+1)B^2.$$

For the other term, just remark that:

$$\left\|\Pi_P^{\mathcal{F},N}\hat{f}-f\right\|_P^2=\left\|\Pi_P^{\mathcal{F},m}\Pi_P^{m,\varepsilon}...\Pi_P^{1,\varepsilon}0-f\right\|_P^2\leq\left\|\Pi_P^{\mathcal{F},m}\Pi_P^{m',\varepsilon}...\Pi_P^{1,\varepsilon}0-f\right\|_P^2$$
$$=\left\|\Pi_P^{\mathcal{F},m'}\Pi_P^{m',\varepsilon}...\Pi_P^{1,\varepsilon}0-f\right\|_P^2$$
$$\leq\sum_{k=1}^{m'}\frac{4\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\left[\frac{1}{N}\sum_{i=1}^{N}f_{\theta_k}(X_i)^2Y_i^2+B^2+\sigma^2\right]+\left\|\overline{\theta}_{m'}-f\right\|_P^2.$$

This is where theorem 2.4 has been used as an oracle inequality: the estimator that we have, with $m\geq m'$, is better than the one with the "good choice" m'. We have too:

$$P^{\otimes N}\left[\mathbb{1}_{\mathcal{E}(\varepsilon)}\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2\right]$$
$$\leq P^{\otimes N}\left[\sum_{k=1}^{m'}\frac{4\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\left[\frac{1}{N}\sum_{i=1}^{N}f_{\theta_k}(X_i)^2Y_i^2+B^2+\sigma^2\right]\right]+(m')^{-2\beta}C$$
$$\leq m'\frac{8\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\left[B^2+\sigma^2\right]$$

So finally, we obtain, for any $m'\leq m$:

$$P^{\otimes N}\left[\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2\right]\leq m'\frac{8\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\left[B^2+\sigma^2\right]$$
$$+(m')^{-2\beta}C+2\varepsilon(m+1)B^2.$$

The choice of:

$$m'=N^{\frac{1}{2\beta+1}}$$

leads to a first term of order $N^{\frac{-2\beta}{2\beta+1}}\log\frac{m}{\varepsilon}$ and a second term of order $N^{\frac{-2\beta}{2\beta+1}}$. The choice of $m=N$ and $\varepsilon=N^{-2}$ gives a first term of order $N^{\frac{-2\beta}{2\beta+1}}\log N$ while keeping the second term at order $N^{\frac{-2\beta}{2\beta+1}}$ and the last term at order N^{-1}. This proves the theorem. □

References

[1] G. Blanchard, P. Massart, R. Vert and L. Zwald, Kernel Projection Machine: a New Tool for Pattern Recognition, *Proceedings of NIPS* 2004.

[2] B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144-152. ACM Press, 1992.

[3] L. Birg, An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics*, 113-133, Leiden, 1999.

[4] O. Catoni, Statistical learning theory and stochastic optimization, *Lecture notes, Saint-Flour summer school on Probability Theory, 2001*, Springer, to appear.

[5] O. Catoni, A PAC-Bayesian approach to adaptative classification, *preprint Laboratoire de Probabilits et Modles Alatoires 2003*.

[6] O. Catoni, Improved Vapnik Cervonenkis bounds, *preprint Laboratoire de Probabilits et Modles Alatoires 2005*.

[7] N. Cristianini and J. Shawe Taylor, *An introduction to Support Vector Machines and other kernel based learning methods,* Cambridge University Press, 2000.

[8] D. L. Donoho and I. M. Johnstone, Ideal Spatial Adaptation by Wavelets, *Biometrika*, Vol. 81, No. 3 (Aug., 1994), 425-455.

[9] G. Kerkyacharian and D. Picard, Regression in random desgin and warped wavelets, *preprint Laboratoire de Probabilits et Modles alatoires 2003*

[10] D. Panchenko, Symmetrization Approach to Concentration Inequalities for Empirical Processes, *The Annals Of Probability*, Vol. 31, No. 4 (2003), 2068-2081.

[11] R Development Core Team, R: A Language And Environment For Statistical Computing, *R Foundation For Statistical Computing*, Vienna, Austria, 2004. URL http://www.R-project.org.

[12] V. N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 1998.

[13] B. Schlkopf, A. J. Smola and K.-R. Mller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10:1299:1319, 1998.

[14] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Journal of Machine Learning Research* **3** (2002), 233–269.

[15] B. Widrow and M. Hoff, *Adaptative swicthing circuits*, IRE WESCON Convention Record, 4:96-104, 1960.

Laboratoire de Probabilits et Modles Alatoires, Universit Paris 6, and Laboratoire de Statistique, CREST, 3, avenue Pierre Larousse, 92240 Malakoff, France.
URL: http://www.crest.fr/pageperso/alquier/alquier.htm
E-mail address: alquier@ensae.fr